1 Astrostatistics

1.1 The Nature and Scope of Astrostatistics

Astrostatistics is at the same time one of the oldest disciplines, and one of the youngest. The Ionian Greek philosopher Thales of Miletus is credited with correctly predicting a total solar eclipse in central Lydia, which he had claimed would occur in May of 585 BCE. He based this prediction on an examination of records maintained by priests throughout the Mediterranean and Near East. The fact that his prediction was apparently well known, and the fact that the Lydians were engaged in a war with the Medes in Central Lydia during this period, brought his prediction notice and fame. Thales was forever after regarded as a sage and even today he is named the father of philosophy and the father of science in books dealing with these subjects.

Thales' success spurred on others to look for natural relationships governing the motions of astronomical bodies. Of particular note was Hipparchus (190–120 BCE) who, following on the earlier work of Aristarchus of Samos (310–230 BCE and Eratosthenes (276–147 BCE), is widely regarded as the first to clearly apply statistical principles to the analysis of astronomical events. Hipparchus also is acknowledged to have first developed trigonometry, spherical trigonometry, and trigonometric tables, applying these to the motions of both the moon and sun. Using the size of the moon's parallax and other data from the median percent of the Sun covered by the shadow of the Earth at various sites in the area, he calculated the distance from the Earth to the Moon as well as from the Earth to the Sun in terms of the Earth's radius. His result was that the median value is 60.5 Earth radii. The true value is 60.3. He also calculated the length of the topical year to within six minutes per year of its true value.

Others in the ancient world, as well as scientists until the early nineteenth century, also used descriptive statistical techniques to describe and calculate the movements and relationships between the Earth and astronomical bodies. Even the first application of a normal or ordinary least squares regression was to astronomy. In 1801 Hungarian Franz von Zach applied the new least squares regression algorithm developed by Carl Gauss for predicting the position of Ceres as it came into view from its orbit behind the Sun.

The development of the first inferential statistical algorithm by Gauss, and its successful application by von Zach, did not immediately lead to major advances in inferential statistics. Astronomers by and large seemed satisfied to work with the Gaussian, or normal, model for predicting astronomical events, and statisticians turned much of their attention to

deriving various probability functions. In the early twentieth century William Gosset, Karl Pearson, and Ronald Fisher made the most advances in statistical modeling and hypothesis testing. Pearson developed the mathematics of goodness-of-fit and of hypothesis testing. The Pearson χ^2 test is still used as an assessment of frequentist based model fit.[1] In addition, Pearson developed a number of tests related to correlation analysis, which is important in both frequentist and Bayesian modeling. Fisher (1890–1962) is widely regarded as the father of modern statistics. He is basically responsible for the frequentist interpretation of hypothesis testing and of statistical modeling. He developed the theories of maximum likelihood estimation and analysis of variance and the standard way that statisticians understood p-values and confidence intervals until the past 20 years. Frequentists still employ his definition of the p-value in their research. His influence on twentieth century statistics cannot be overestimated.

It is not commonly known that Pierre-Simon Laplace (1749–1827) is the person foremost responsible for bringing attention to the notion of Bayesian analysis, which at the time meant employing inverse probability to the analysis of various problems. We shall discuss Bayesian methodology in a bit more detail in Chapter 3. Here we can mention that Thomas Bayes (1702–1761) developed the notion of inverse probability in unpublished notes he made during his lifetime. These notes were discovered by Richard Price, Bayes's literary executor and a mathematician, who restructured and presented Bayes' paper to the Royal Society. It had little impact on British mathematicians at the time, but it did catch the attention of Laplace. Laplace fashioned Bayes' work into a full approach to probability, which underlies current Bayesian statistical modeling. It would probably be more accurate to call Bayesian methodology Bayes–Laplace methodology, but simplification has given Bayes the nominal credit for this approach to both probability and statistical modeling.

After enthusiastically promoting inverse probability, Laplace abandoned this work and returned to researching probability theory from the traditional perspective. He also made major advances in differential equations, including his discovery of Laplace transforms, which are still very useful in mathematics. Only a few adherents to the Bayesian approach to probability carried on the tradition throughout the next century and a half. Mathematicians such as Bruno de Finetti (Italy) and Harold Jeffreys (UK) promoted inverse probability and Bayesian analysis during the early part of the twentieth century, while Dennis Lindley (UK), Leonard J Savage (US), and Edwin Jaynes, a US physicist, were mainstays of the tradition in the early years of the second half of the twentieth century. But their work went largely unnoticed.

The major problem with Bayesian analysis until recent times has been related to the use of priors. Priors are distributions representing information from outside the model data that is incorporated into the model. We shall be discussing priors in some detail later in the text. Briefly though, except for relatively simple models the mathematics of calculating so-called posterior distributions was far too difficult to do by hand, or even by most computers, until computing speed and memory became powerful enough. This was particularly the case for Bayesian models, which are extremely demanding on computer power. The foremost reason why most statisticians and analysts were not interested in

[1] In Chapter 5 we shall discuss how this statistic can be used in Bayesian modeling as well.

implementing Bayesian methods into their research was that computing technology was not advanced enough to execute more than fairly simple problems. This was certainly the case with respect to the use of Bayesian methods in astronomy.

We shall provide a brief overview of Bayesian methodology in Chapter 3. Until such methods became feasible, however, Fisherian or frequentist methodology was the standard way of doing statistics. For those who are interested in understanding the history of this era of mathematics and statistics, we refer you to the book *Willful Ignorance: The Mismeasure of Uncertainty* (Weisberg, 2014).

Astronomy and descriptive statistics, i.e., the mathematics of determining mean, median, mode, range, tabulations, frequency distributions, and so forth, were closely tied together until the early nineteenth century. Astronomers have continued to employ descriptive statistics, as well as basic linear regression, to astronomical data. But they did not concern themselves with the work being done in statistics during most of the twentieth century. Astronomers found advances in telescopes and the new spectroscope, as well as in calculus and differential equations in particular, to be much more suited to understanding astrophysical data than hypothesis testing and other frequentist methods. There was a near schism between astronomers and statisticians until the end of the last century.

As mentioned in the Preface, astrostatistics can be regarded as the statistical analysis of astronomical data. Unlike how astronomers utilized statistical methods in the past, primarily focusing on descriptive measures and to a moderate extent on linear regression from within the frequentist tradition, astrostatistics now entails the use, by a growing number of astronomers, of the most advanced methods of statistical analysis that have been developed by members of the statistical profession. However, we still see astronomers using linear regression on data that to a statistician should clearly be modeled by other, nonlinear, means. Recent texts on the subject have been aimed at informing astronomers of these new advanced statistical methods.

It should be mentioned that we have incorporated astroinformatics under the general rubric of astrostatistics. Astroinformatics is the study of the data gathering and computing technology needed to gather astronomical data. It is essential to statistical analysis. Some have argued that astroinformatics incorporates astrostatistics, which is the reverse of the manner in which we envision it. But the truth is that gathering information without an intent to analyze it is a fairly useless enterprise. Both must be understood together. The International Astronomical Union (IAU) has established a new commission on astroinformatics and astrostatistics, seeming to give primacy to the former, but how the order is given depends somewhat on the interests of those who are establishing such names. Since we are focusing on statistics, although we are also cognizant of the important role that informatics and information sciences bring to bear on statistical analysis, we will refer to the dual studies of astrostatistics and astroinformatics as simply astrostatistics.

In the last few decades the size and complexity of the available astronomical information has closely followed Moore's law, which states roughly that computing processing power doubles every two years. However, our ability to acquire data has already surpassed our capacity to analyze it. The Sloan Digital Sky Survey (SDSS), in operation since 2000, was one of the first surveys to face the big data challenge: in its first data release it observed 53

million individual objects. The Large Synoptic Survey Telescope (LSST) survey is aiming to process and store 30 terabytes of data each night for a period of ten years. Not only must astronomers (astroinformaticists) deal with such massive amounts of data but also it must be stored in a form in which meaningful statistical analysis can proceed. In addition, the data being gathered is permeated with environmental noise (due to clouds, moon, interstellar and intergalactic dust, cosmic rays), instrumental noise (due to distortions from telescope design, detector quantum efficiency, a border effect in non-central pixels, missing data), and observational bias (since, for example, brighter objects have a higher probability of being detected). All these problem areas must be dealt with prior to any attempt to subject the data to statistical analysis. The concerns are immense, but nevertheless this is a task which is being handled by astrostatisticians and information scientists.

1.2 The Recent Development of Astrostatistics

Statistical analysis has developed together with the advance in computing speed and storage. Maximum likelihood and all such regression procedures require that a matrix be inverted. The size of the matrix is based on the number of predictors and parameters in the model, including the intercept. Parameter estimates can be determined using regression for only very small models if done by hand. Beginning in the 1960s, larger regression and multivariate statistical procedures could be executed on mainframe computers using SAS, SPSS, and other statistical and data management software designed specifically for mainframe systems. These software packages were ported to the PC environment when PCs with hard drives became available in 1983. Statistical routines requiring a large number of iterations could take a long time to converge. But as computer speeds became ever faster, new regression procedures could be developed and programmed that allowed for the rapid inversion of large matrices and the solution to complex modeling projects.

Complex Bayesian modeling, if based on Markov chain Monte Carlo (MCMC) sampling, was simply not feasible until near the turn of the century. By 2010 efficient Bayesian sampling algorithms were implemented into statistical software, and the computing power by then available allowed astronomers to analyze complex data situations using advanced statistical techniques. By the closing years of the first decade of this century astronomers could take advantage of the advances in statistical software and of the training and expertise of statisticians.

As mentioned before, from the early to middle years of the nineteenth century until the final decade of the twentieth century there was little communication between astronomers and statisticians. We should note, though, that there were a few astronomers who were interested in applying sophisticated statistical models to their study data. In 1990, for example, Thomas Loredo wrote his PhD dissertation in astrophysics at the University of Chicago on “From Laplace Supernova SN 1987A: Bayesian inference in astrophysics” (see the book Feigelson and Babu, 2012a), which was the first thorough application of Bayesian modeling to astrophysical data. It was, and still is, the seminal work in the area and can be regarded as one of the founding articles in the new discipline of astrostatistics.

In 1991 Eric Feigelson and Jogesh Babu, an astronomer and statistician respectively at Pennsylvania State University, collaboratively initiated a conference entitled *Statistical Challenges in Modern Astronomy*. The conference was held at their home institution and brought together astronomers and a few statisticians for the purpose of collaboration. The goal was also to find a forum to teach astronomers how to use appropriate statistical analysis for their study projects. These conferences were held every five years at Penn State until 2016. The conference site has now shifted to Carnegie Mellon University under the direction of Chad Schafer. During the 1990s and 2000s a few other conferences, workshops, and collaborations were held that aimed to provide statistical education to astronomers. But they were relatively rare, and there did not appear to be much growth in the area.

Until 2008 there were no astrostatistics working groups or committees authorized under the scope of any statistical or astronomical association or society. Astrostatistics was neither recognized by the IAU nor recognized as a discipline in its own right by the principal astronomical and statistical organizations. However, in 2008 the first interest group in astrostatistics was initiated under the International Statistical Institute (ISI), the statistical equivalent of the IAU for astronomy. This interest group, founded by the first author of the book, and some 50 other interested astronomers and statisticians, met together at the 2009 ISI World Statistics Congress in Durban, South Africa. The attendees voted to apply for the creation of a standing committee on astrostatistics under the auspices of the ISI. The proposal was approved at the December 2009 meeting of the ISI executive board. It was the first such astrostatistics committee authorized by an astronomical or statistical society.

In the following month the ISI committee expanded to become the ISI Astrostatistics Network. Network members organized and presented two invited sessions and two special theme sessions in astrostatistics at the 2011 World Statistics Congress in Dublin, Ireland. Then, in 2012, the International Astrostatistics Association (IAA) was formed from the Network as an independent professional scientific association for the discipline. Also in 2012 the Astrostatistics and Astroinformatics Portal (ASAIP) was instituted at Pennsylvania State under the editorship of Eric Feigelson and the first author of this volume. As of January 2016 the IAA had over 550 members from 56 nations, and the Portal had some 900 members. Working Groups in astrostatistics and astroinformatics began under the scope of IAU, the American Astronomical Society, and the American Statistical Association. In 2014 the IAA sponsored the creation of its first section, the Cosmostatistics Initiative (COIN), led by the second author of this volume. In 2015 the IAU working group became the IAU Commission on Astroinformatics and Astrostatistics, with Feigelson as its initial president. Astrostatistics, and astroinformatics, is now a fully recognized discipline. Springer has a series on astrostatistics, the IAA has begun agreements with corporate partnerships, Cambridge University Press is sponsoring IAA awards for contributions to the discipline, and multiple conferences in both astrostatistics and astroinformatices are being held – all to advance the discipline. The IAA headquarters is now at Brera Observatory in Milan. The IAA website is located at: http://iaa.mi.oa-brera.inaf.it and the ASAIP Portal URL is https://asaip.psu.edu. We recommend the readers of this volume to visit these websites for additional resources on astrostatistics.

1.3 What is a Statistical Model?

Statistics and statistical modeling have been defined in a variety of ways. We shall define it in a general manner as follows:

> *Statistics may be generically understood as the science of collecting and analyzing data for the purpose of classification, prediction, and of attempting to quantify and understand the uncertainty inherent in phenomena underlying data.*
>
> (J.M. Hilbe, 2014)

Note that this definition ties data collection and analysis under the scope of statistics. This is analogical to how we view astrostatistics and astroinformatics. In this text our foremost interest is in parametric models. Since the book aims to develop code and to discuss the characterization of a number of models, it may be wise to define what is meant by statistics and statistic models. Given that many self-described data scientists assert that they are not statisticians and that statistics is dying, we should be clear about the meaning of these terms and activities.

Statistical models are based on probability distributions. Parametric models are derived from distributions having parameters which are estimated in the execution of a statistical model. Non-parametric models are based on empirical distributions and follow the natural or empirical shape of the data. We are foremost interested in parametric models in this text.

The theoretical basis of a statistical model differs somewhat from how analysts view a model when it is estimated and interpreted. This is primarily the case when dealing with models from a frequentist-based view. In the frequentist tradition, the idea is that the data to be modeled is generated by an underlying probability distribution. The researcher typically does not observe the entire population of the data that is being modeled but rather observes a random sample from the population data, which itself derives from a probability distribution with specific but unknown fixed parameters. The parameters specifying both the population and sample data consist of the distributional mean and one or more shape or scale parameters. The mean is regarded as a location parameter. For the normal model the variance σ^2 is the scale parameter, which has a constant value across the range of observations in the data. Binomial and count models have a scale parameter but its value is set at unity. The negative binomial has a dispersion parameter, but it is not strictly speaking a scale parameter. It adjusts the model for extra correlation or dispersion in the data. We shall address these parameters as we discuss various models in the text.

In frequentist-based modeling the slopes or coefficients that are derived in the estimation process for explanatory predictors are also considered as parameters. The aim of modeling is to estimate the parameters defining the probability distribution that is considered to generate the data being modeled. The predictor coefficients, and intercept, are components of the mean parameter.

In the Bayesian tradition parameters are considered as randomly distributed, not as fixed. The data is characterized by an underling probability distribution but each parameter is separately estimated. The distribution that is used to explain the predictor and parameter data is called the likelihood. The likelihood may be mixed with outside or additional

information known from other studies or obtained from the experience or background of the analyst. This external information, when cast as a probability distribution with specified parameters, is called the prior distribution. The product of the model (data) likelihood and prior distributions is referred to as the posterior distribution. When the model is simple, the posterior distribution of each parameter may be analytically calculated. However, for most real model data, and certainly for astronomical data, the posterior must be determined through the use of a sampling algorithm. A variety of MCMC sampling algorithms or some version of Gibbs sampling are considered to be the standard sampling algorithms used in Bayesian modeling. The details of these methods go beyond the scope of this text. For those interested in sampling algorithms we refer you to the articles Feroz and Hobson (2008) and Foreman-Mackey *et al.* (2013) and the books Gamerman and Lopes (2006), Hilbe and Robinson (2013), and Suess and Trumbo (2010).

1.4 Classification of Statistical Models

We mentioned earlier that statistical models are of two general varieties – parametric and non-parametric. Parametric models are based on a probability distribution, or a mixture of distributions. This is generally the case for both frequentist-based and Bayesian models. Parametric models are classified by the type of probability distribution upon which a model is based. In Figure 1.1 we provide a non-exhaustive classification of the major models discussed in this volume.

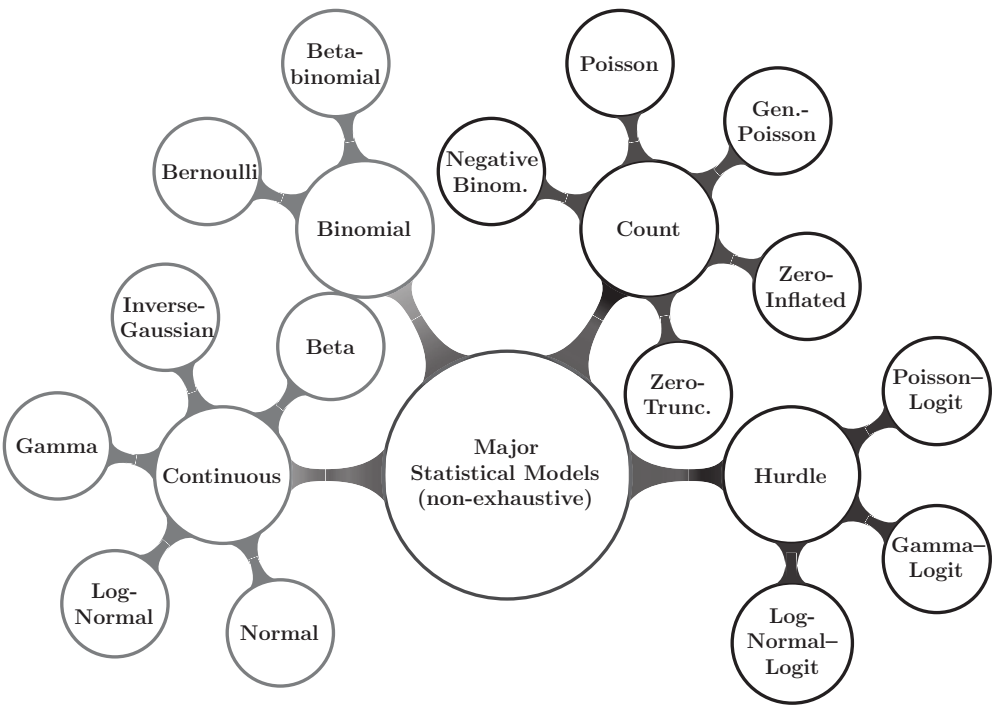

Figure 1.1 Major classifications of statistical regression models.

Astronomers utilize, or should utilize, most of these model types in their research. In this text examples will be given, and code provided, for all. We aim to show how only relatively minor changes need to be given to the basic estimation algorithm in order to develop and expand the models presented in Figure 1.1. Regarding data science and statistics, it is clear that when one is engaged in employing statistical models to evaluate data – whether for better understanding of the data or to predict observations beyond it – one is doing statistics. Statistics is a general term characterizing both descriptive and predictive measures of data and represents an attempt to quantify the uncertainty inherent in both the data being evaluated and in our measuring and modeling mechanisms. Many statistical tools employed by data scientists can be of use in astrostatistics, and we encourage astronomers and traditional statisticians to explore how they can be used to obtain a better evaluation of astronomical data. Our focus on Bayesian modeling can significantly enhance this process.

Although we could have begun our examination of Bayesian models with single-parameter Bernoulli-based models such as Bayesian logistic and probit regression, we shall first look at the normal or Gaussian model. The normal distribution has two characteristic parameters, the mean or location parameter and the variance or scale parameter. The normal model is intrinsically more complex than single-parameter models, but it is the model most commonly used in statistics – from both the frequentist and Bayesian traditions. We believe that most readers will be more familiar with the normal model from the outset and will have worked with it in the past. It therefore makes sense to begin with it. The Bayesian logistic and probit binary response models, as well as the Poisson count model, are, however, easier to understand and work with than the normal model.

Further Reading

Feigelson, E. D. and J. G. Babu (2012). *Statistical Challenges in Modern Astronomy V.* Lecture Notes in Statistics. Springer.

Hilbe, J. M. (2012). "Astrostatistics in the international arena." *In: Statistical Challenges in Modern Astronomy V,* eds. E. D. Feigelson and J. G. Babu. Springer, pp. 427–433.

Hilbe, J. M. (2016). "Astrostatistics as new statistical discipline – a historical perspective." `www.worldofstatistics.org/files/2016/05/WOS_newsletter_05252016.pdf` (visited on 06/16/2016).

McCullagh, P. (2002). "What is a statistical model?" *Ann. Statist.* 30(5), 1225–1310. DOI: `10.1214/aos/1035844977`.

White, L. A. (2014). "The rise of astrostatistics." `www.symmetrymagazine.org/article/november-2014/the-rise-of-astrostatistics` (visited on 06/16/2016).