A8824: Reader's Guide to Ivezic et al. Chapters 3-5

Chapter 3

Section 3.1 introduces the basics of probability and random variables, and it's all worth reading; don't spend too much time parsing through the contingency table.

Section 3.2 describes basic statistical summaries of distributions. This should be familiar material, easily skimmed. The notion of interquartile range as a robust estimator for the width of a distribution (equation 3.36) is a useful one.

Section 3.3 runs through the large zoo of standard univariate probability distribution functions. All of these are useful in some circumstance or another, but the ones you should pay the most attention to are: the Gaussian distribution (3.3.2) the Poisson distribution (3.3.4), the Cauchy distribution (3.3.5) as the most common example of a "pathological case," and the χ^2 distribution (3.3.7), which plays an important role in error analysis.

Section 3.4 summarizes the famed (and important) central limit theorem, that a random variable that is the sum of many independent random variables will have a Gaussian distribution.

Section 3.5 turns to bivariate and multivariate distributions. You should pay attention to the basic definitions in 3.5.1, the bivariate Gaussian distribution in 3.5.2, and its generalization to a multivariate Gaussian in 3.5.4. These are frequently needed to model data with correlated errors.

Section 3.6 defines a variety of correlation coefficients and tests for whether distributions are correlated. You can skim this material and come back to it when you run into a reason to learn the details of the Pearson, Spearman or Kendall coefficient. These are useful basic tests.

Section 3.7 describes a useful general method for generating random numbers drawn from an arbitrary distribution, which you may often find yourself needing to do.

Chapter 4

Chapter 4.1 summarizes the difference of classical (a.k.a. frequentist) statistics and Bayesian statistics, which will be returned to in Chapter 5.

Chapter 4.2 addresses the important general topic of maximum likelihood estimation. It plays a key role in parameter estimation problems in classical statistics, and it is carried over with some modification in Bayesian statistics. This section is worth close attention.

Section 4.3.1, on goodness-of-fit, explains the important rules of thumb for evaluating the significance of χ^2 values. For model comparison (4.3.2), I generally prefer a Bayesian approach, but if you run into the term Aikake Information Criterion and want to know what it means, this is where you find it.

Section 4.4, on Gaussian mxtures and the EM algorithm, is a relatively advanced topic and can be skipped or skimmed on first reading. It is interesting for automated classification problems (a major theme of this book).

Section 4.5 describes two very useful practical tools for estimating confidence intervals, bootstrap and jackknife.

Section 4.6 describes the classical approach to hypothesis testing, and the issue of false positives. It's worth skimming, but I generally prefer a Bayesian approach to this class of question.

Section 4.7 discusses the specific case of assessing the hypothesis that two samples are drawn from the same distribution (or one sample is drawn from a specified distribution). The Kolmogorov-

Smirnov (K-S) test described in 4.7.2 is widely used and worth understanding. The rest of this section, describing a variety of other tests for specific cases, can be skimmed.

Section 4.8 gives some quite useful advice on choosing bin widths and estimating errors for histograms.

Section 4.9 addresses a more specialized topic of assessing a luminosity function in the presence of selection effects; it's a good starting point if you have to solve a problem of this sort.

Chapter 5

Section 5.1 gives a good overview and summary of Bayesian statistical inference.

Section 5.2 discusses methods for choosing priors. It's a bit technical, with §5.2.1 being the most useful.

Section 5.3 (short) introduces the concepts of credible regions (analogous to frequentist confidence intervals) and marginalization over nuisance parameters, both important in parameter estimation problems.

Section 5.4 discusses Bayesian model selection/hypothesis comparison. The introduction and §5.4.1 introduce the general concept. Section 5.4.2 gives a nice example that illustrates how Bayesian hypothesis comparison automatically incorporates "Occam's razor," preferring the simplest hypothesis consistent with the data. Section 5.4.3 describes the Bayeian Information Criterion, which is an approximate method of estimating an odds ratio of models when a full calculation may be impractical.

Section 5.5 discusses biases that arise when the underlying distribution from which a data sample is drawn is non-uniform. These biases are frequently important in astronomical applications and a common cause of confusion. They can be understood in a nicely unified way in a Bayesian context.

Section 5.6 gives a number of examples of Bayesian parameter estimation. While worked examples are useful, this section is rather long and detailed. Section 5.6.1 illustrates useful general points. Section 5.6.5 is a useful example of marginalization in the context of detecting a signal in the presence of background. Section 5.6.6 discusses straight-line fitting with Poisson vs. Gaussian errors, which is relevant to some problems. Section 5.6.7 discusses an interesting application of Bayesian methodology where one can reject "outlier" data points in a statistically sensible way.

Section 5.7 gives examples of Bayesian model selection, including application to the problem of constructing optimal (variable width) histogram bins discussed previously in §4.8.

Section 5.8 discusses MCMC methods, with a useful general discussion and an introduction to python tools for MCMC. §§5.8.5 and 5.8.6 give some worked examples.

Section 5.9 is a very good, balanced summary of the relative strengths and shortcomings of frequentist and Bayesian statistics. I think it's consistent with my advice: one should be a Bayesian in principle, but not always in practice.