CHAPTER

13 Bayesian Estimation in Hierarchical Models

John K. Kruschke and Wolf Vanpaemel

Kruschke, J. K. and Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. In: J. R. Busemeyer, Z. Wang, J. T. Townsend, and A. Eidels (Eds.), The Oxford Handbook of Computational and Mathematical Psychology, pp. 279-299. Oxford, UK: Oxford University Press.

Abstract

Bayesian data analysis involves describing data by meaningful mathematical models, and allocating credibility to parameter values that are consistent with the data and with prior knowledge. The Bayesian approach is ideally suited for constructing hierarchical models, which are useful for data structures with multiple levels, such as data from individuals who are members of groups which in turn are in higher-level organizations. Hierarchical models have parameters that meaningfully describe the data at their multiple levels and connect information within and across levels. Bayesian methods are very flexible and straightforward for estimating parameters of complex hierarchical models (and simpler models too). We provide an introduction to the ideas of hierarchical models and to the Bayesian estimation of their parameters, illustrated with two extended examples. One example considers baseball batting averages of individual players grouped by fielding position. A second example uses a hierarchical extension of a cognitive process model to examine individual differences in attention allocation of people who have eating disorders. We conclude by discussing Bayesian model comparison as a case of hierarchical modeling.

Key Words: Bayesian statistics, Bayesian data analysis, Bayesian modeling, hierarchical model, model comparison, Markov chain Monte Carlo, shrinkage of estimates, multiple comparisons, individual differences, cognitive psychometrics, attention allocation

The Ideas of Hierarchical Bayesian Estimation

Bayesian reasoning formalizes the reallocation of credibility over possibilities in consideration of new data. Bayesian reasoning occurs routinely in everyday life. Consider the logic of the fictional detective Sherlock Holmes, who famously said that when a person has eliminated the impossible, then whatever remains, no matter how improbable, must be the truth (Doyle, 1890). His reasoning began with a set of candidate possibilities, some of which had low credibility *a priori*. Then he collected evidence through detective work, which ruled out some possibilities. Logically, he then reallocated credibility to the remaining possibilities. The complementary logic of judicial exoneration is also commonplace. Suppose there are several unaffiliated suspects for a crime. If evidence implicates one of them, then the other suspects are exonerated. Thus, the initial allocation of credibility (i.e., culpability) across the suspects was reallocated in response to new data.

In data analysis, the space of possibilities consists of parameter values in a descriptive model. For example, consider a set of data measured on a continuous scale, such as the weights of a group of 10-year-old children. We might want to describe the set of data in terms of a mathematical normal distribution, which has two parameters, namely the mean and the standard deviation. Before collecting the data, the possible means and standard deviations have some prior credibility, about which

we might be very uncertain or highly informed. After collecting the data, we reallocate credibility to values of the mean and standard deviation that are reasonably consistent with the data and with our prior beliefs. The reallocated credibilities constitute the posterior distribution over the parameter values.

We care about parameter values in formal models because the parameter values carry meaning. When we say that the mean weight is 32 kilograms and the standard deviation is 3.2 kilograms, we have a clear sense of how the data are distributed (according to the model). As another example, suppose we want to describe children's growth with a simple linear function, which has a slope parameter. When we say that the slope is 5 kilograms per year, we have a clear sense of how weight changes through time (according to the model). The central goal of Bayesian estimation, and a major goal of data analysis generally, is deriving the most credible parameter values for a chosen descriptive model, because the parameter values are meaningful in the context of the model.

Bayesian estimation provides an entire distribution of credibility over the space of parameter values, not merely a single "best" value. The distribution precisely captures our uncertainty about the parameter estimate. The essence of Bayesian estimation is to formally describe how uncertainty changes when new data are taken into account.

Hierarchical Models Have Parameters with Hierarchical Meaning

In many situations, the parameters of a model have meaningful dependencies on each other. As a simplistic example, suppose we want to estimate the probability that a type of trick coin, manufactured by the Acme Toy Company, comes up heads. We know that different coins of that type have somewhat different underlying biases to come up heads, but there is a central tendency in the bias imposed by the manufacturing process. Thus, when we flip several coins of that type, each several times, we can estimate the underlying biases in each coin and the typical bias and consistency of the manufacturing process. In this situation, the observed heads of a coin depend only on the bias in the individual coin, but the bias in the coin depends on the manufacturing parameters. This chain of dependencies among parameters exemplifies a hierarchical model (Kruschke, 2015, Ch. 9).

As another example, consider research into childhood obesity. The researchers measure weights of children in a number of different schools that have different school lunch programs, and from a number of different school districts that may have different but unknown socioeconomic statuses. In this case, a child's weight might be modeled as dependent on his or her school lunch program. The school lunch program is characterized by parameters that indicate the central tendency and variability of weights that it tends to produce. The parameters of the school lunch program are, in turn, dependent on the school's district, which is described by parameters indicating the central tendency and variability of school-lunch parameters across schools in the district. This chain of dependencies among parameters again exemplifies a hierarchical model.

In general, a model is hierarchical if the probability of one parameter can be conceived to depend on the value of another parameter. Expressed formally, suppose the observed data, denoted D, are described by a model with two parameters, denoted α and β. The probability of the data is a mathematical function of the parameter values, denoted by $p(D|\alpha,\beta)$, which is called the *likelihood* function of the parameters. The prior probability of the parameters is denoted $p(\alpha,\beta)$. Notice that the likelihood and prior are expressed, so far, in terms of combinations of α and β in the joint parameter space. The probability of the data, weighted by the probability of the parameter values, is the product, $p(D|\alpha,\beta)p(\alpha,\beta)$. The model is *hierarchical* if that product can be factored as a chain of dependencies among parameters, such as $p(D|\alpha,\beta)p(\alpha,\beta) = p(D|\alpha)p(\alpha|\beta)p(\beta)$.

Many models can be reparameterized, and conditional dependencies can be revealed or obscured under different parameterizations. The notion of hierarchical has to do with a particular meaningful definition of a model structure that expresses dependencies among parameters in a meaningful way. In other words, it is the *semantics* of the parameters when factored in the corresponding way that makes a model hierarchical. Ultimately, any multiparameter model merely has parameters in a joint space, whether that joint space is *conceived* as hierarchical or not. Many realistic situations involve natural hierarchical meaning, as illustrated by the two major examples that will be described at length in this chapter.

One of the primary applications of hierarchical models is describing data from individuals within

Kruschke, J. K. and Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. In: J. R. Busemeyer, Z. Wang, J. T. Townsend, and A. Eidels (Eds.), The Oxford Handbook of Computational and Mathematical Psychology, pp. 279-299. Oxford, UK: Oxford University Press.

groups. A hierarchical model may have parameters for each individual that describe each individual's tendencies, and the distribution of individual parameters within a group is modeled by a higher-level distribution with its own parameters that describe the tendency of the group. The individual-level and group-level parameters are estimated simultaneously. Therefore, the estimate of each individual-level parameter is informed by all the other individuals via the estimate of the group-level distribution, and the group-level parameters are more precisely estimated by the jointly constrained individual-level parameters. The hierarchical approach is better than treating each individual independently because the data from different individuals meaningfully inform one another. And the hierarchical approach is better than collapsing all the individual data together because collapsed data may blur or obscure trends within each individual.

Advantages of the Bayesian Approach

Bayesian methods provide tremendous flexibility in designing models that are appropriate for describing the data at hand, and Bayesian methods provide a complete representation of parameter uncertainty (i.e., the posterior distribution) that can be directly interpreted. Unlike the frequentist interpretation of parameters, there is no construction of sampling distributions from auxiliary null hypotheses. In a frequentist approach, although it may be possible to find a maximum-likelihood estimate (MLE) of parameter values in a hierarchical nonlinear model, the subsequent task of interpreting the uncertainty of the MLE can be very difficult. To decide whether an estimated parameter value is significantly different from a null value, frequentist methods demand construction of sampling distributions of arbitrarily-defined deviation statistics, generated from arbitrarily-defined null hypotheses, from which p values are determined for testing null hypotheses. When there are multiple tests, frequentist decision rules must adjust the p values. Moreover, frequentist methods are unwieldy for constructing confidence intervals on parameters, especially for complex hierarchical nonlinear models that are often the primary interest for cognitive scientists.[1] Furthermore, confidence intervals change when the researcher intention changes (e.g., Kruschke, 2013). Frequentist methods for measuring uncertainty (as confidence intervals from sampling distributions) are fickle and difficult, whereas Bayesian methods are inherently designed to provide clear representations of uncertainty. A thorough critique of frequentist methods such as p values would take us too far afield. Interested readers may consult many other references, such as articles by Kruschke (2013) or Wagenmakers (2007).

Some Mathematics and Mechanics of Bayesian Estimation

The mathematically correct reallocation of credibility over parameter values is specified by Bayes' rule (Bayes & Price, 1763):

$$\underbrace{p(\alpha|D)}_{\text{posterior}} = \underbrace{p(D|\alpha)}_{\text{likelihood}} \underbrace{p(\alpha)}_{\text{prior}} / p(D) \tag{1}$$

where

$$p(D) = \int d\alpha \, p(D|\alpha) p(\alpha) \tag{2}$$

is called the "marginal likelihood" or "evidence." The formula in Eq. 1 is a simple consequence of the definition of conditional probability (e.g., Kruschke, 2015), but it has huge ramifications when applied to meaningful, complex models.

In some simple situations, the mathematical form of the posterior distribution can be analytically derived. These cases demand that the integral in Eq. 2 can be mathematically derived in conjunction with the product of terms in the numerator of Bayes' rule. When this can be done, the result can be especially pleasing because an explicit, simple formula for the posterior distribution is obtained.

Analytical solutions for Bayes' rule can rarely be achieved for realistically complex models. Fortunately, instead, the posterior distribution is approximated, to arbitrarily high accuracy, by generating a huge random sample of representative parameter values from the posterior distribution. A large class of algorithms for generating a representative random sample from a distribution is called Markov chain Monte Carlo (MCMC) methods. Regardless of which particular sampler from the class is used, in the long run they all converge to an accurate representation of the posterior distribution. The bigger the MCMC sample, the finer-resolution picture we have of the posterior distribution. Because the sampling process uses a Markov chain, the random sample produced by the MCMC process is often called a chain.

Kruschke, J. K. and Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. In: J. R. Busemeyer, Z. Wang, J. T. Townsend, and A. Eidels (Eds.), The Oxford Handbook of Computational and Mathematical Psychology, pp. 279-299. Oxford, UK: Oxford University Press.

Box 1 MCMC Details

Because the MCMC sampling is a random walk through parameter space, we would like some assurance that it successfully explored the posterior distribution without getting stuck, oversampling, or undersampling zones of the posterior. Mathematically, the samplers will be accurate in the long run, but we do not know in advance exactly how long is long enough to produce a reasonably good sample.

There are various diagnostics for assessing MCMC chains. It is beyond the scope of this chapter to review their details, but the ideas are straightforward. One type of diagnostic assesses how "clumpy" the chain is, by using a descriptive statistic called the autocorrelation of the chain. If a chain is strongly autocorrelated, successive steps in the chain are near each other, thereby producing a clumpy chain that takes a long time to smooth out. We want a smooth sample to be sure that the posterior distribution is accurately represented in all regions of the parameter space. To achieve stable estimates of the tails of the posterior distribution, one heuristic is that we need about 10,000 independent representative parameter values (Kruschke, 2015, Section 7.5.2). Stable estimates of central tendencies can be achieved by smaller numbers of independent values. A statistic called the *effective sample size* (ESS) takes into account the autocorrelation of the chain and suggests what would be an equivalently sized sample of independent values.

Another diagnostic assesses whether the MCMC chain has gotten stuck in a subset of the posterior distribution, rather than exploring the entire posterior parameter space. This diagnostic takes advantage of running two or more distinct chains, and assessing the extent to which the chains overlap. If several different chains thoroughly overlap, we have evidence that the MCMC samples have converged to a representative sample.

It is important to understand that the MCMC "sample" or "chain" is a huge representative sample of parameter values from the posterior distribution. The MCMC sample is not to be confused with the sample of data. For any particular analysis, there is a single fixed sample of data, and there is a single underlying mathematical posterior distribution that is inferred from the sample of data. The MCMC chain typically uses tens of thousands of representative parameter values from the posterior distribution to represent the posterior distribution. Box 1 provides more details about assessing when an MCMC chain is a good representation of the underlying posterior distribution.

Contemporary MCMC software works seamlessly for complex hierarchical models involving nonlinear relationships between variables and non-normal distributions at multiple levels. Model-specification languages such as BUGS (Lunn, Jackson, Best, Thomas, & Spiegelhalter, 2013; Lunn, Thomas, Best, & Spiegelhalter, 2000), JAGS (Plummer, 2003), and Stan (Stan, 2013) allow the user to specify descriptive models to satisfy theoretical and empirical demands.

Example: Shrinkage and Multiple Comparisons of Baseball Batting Abilities

American baseball is a sport in which one person, called a pitcher, throws a small ball as quickly as possible over a small patch of earth, called home plate, next to which is standing another person holding a stick, called a bat, who tries to hit the ball with the bat. If the ball is hit appropriately into the field, the batter attempts to run to other marked patches of earth arranged in a diamond shape. The batter tries to arrive at the first patch of earth, called first base, before the other players, called fielders, can retrieve the ball and throw it to a teammate attending first base.

One of the crucial abilities of baseball players is, therefore, the ability to hit a very fast ball (sometimes thrown more than 90 miles [145 kilometers] per hour) with the bat. An important goal for enthusiasts of baseball is estimating each player's ability to bat the ball. Ability can not be assessed directly but can only be estimated by observing how many times a player was able to hit the ball in all his opportunities at bat, or by observing hits and at-bats from other similar players.

There are nine players in the field at once, who specialize in different positions. These include the pitcher, the catcher, the first base man, the second base man, the third base man, the shortstop, the left fielder, the center fielder, and the right fielder. When one team is in the field, the other team is at bat. The teams alternate being at bat and being in the field. Under some rules, the pitcher does not have to bat when his team is at bat.

Because different positions emphasize different skills while on the field, not all players are prized

Kruschke, J. K. and Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. In: J. R. Busemeyer, Z. Wang, J. T. Townsend, and A. Eidels (Eds.), The Oxford Handbook of Computational and Mathematical Psychology, pp. 279-299. Oxford, UK: Oxford University Press.

for their batting ability alone. In particular, pitchers and catchers have specialized skills that are crucial for team success. Therefore, based on the structure of the game, we know that players with different primary positions are likely to have different batting abilities.

The Data

The data consist of records from 948 players in the 2012 regular season of Major League Baseball who had at least one at-bat.[2] For player i, we have his number of opportunities at bat, AB_i, his number of hits H_i, and his primary position when in the field $pp(i)$. In the data, there were 324 pitchers with a median of 4.0 at-bats, 103 catchers with a median of 170.0 at-bats, and 60 right fielders with a median of 340.5 at-bats, along with 461 players in six other positions.

The Descriptive Model with Its Meaningful Parameters

We want to estimate, for each player, his underlying probability θ_i of hitting the ball when at bat. The primary data to inform our estimate of θ_i are the player's number of hits, H_i, and his number of opportunities at bat, AB_i. But the estimate will also be informed by our knowledge of the player's primary position, $pp(i)$, and by the data from all the other players (i.e., their hits, at-bats, and positions). For example, if we know that player i is a pitcher, and we know that pitchers tend to have θ values around 0.13 (because of all the other data), then our estimate of θ_i should be anchored near 0.13 and adjusted by the specific hits and at-bats of the individual player. We will construct a hierarchical model that rationally shares information across players within positions, and across positions within all major league players.[3]

We denote the i^{th} player's underlying probability of getting a hit as θ_i. (See Box 2 for discussion of assumptions in modeling.) Then the number of hits H_i out of AB_i at-bats is a random draw from a binomial distribution that has success rate θ_i, as illustrated at the bottom of Figure 13.1. The arrow pointing to H_i is labeled with a "$\sim$" symbol to indicate that the number of hits is a random variable distributed as a binomial distribution.

To formally express our prior belief that different primary positions emphasize different skills and hence have different batting abilities, we assume that the player abilities θ_i come from distributions specific to each position. Thus, the θ_i's for the 324 pitchers are assumed to come from a distribution specific to pitchers, that might have a different central tendency and dispersion than the distribution of abilities for the 103 catchers, and so on for the other positions. We model the distribution of θ_i's for a position as a beta distribution, which is a natural distribution for describing values that fall between zero and one, and is often used in this sort of application (e.g., Kruschke, 2015). The mean of the beta distribution for primary position pp is denoted μ_{pp}, and the narrowness of the distribution is denoted κ_{pp}. The value of μ_{pp} represents the typical batting ability of players in primary position pp, and the value of κ_{pp} represents how tightly clustered the abilities are across players in primary position pp. The κ parameter is sometimes called the *concentration* or *precision* of the beta distribution.[4] Thus, an individual player whose primary position is $pp(i)$ is assumed to have a batting ability θ_i that comes from a beta distribution with mean $\mu_{pp(i)}$ and precision $\kappa_{pp(i)}$. The values of μ_{pp} and κ_{pp} are estimated simultaneously with all the θ_i. Figure 13.1 illustrates this aspect of the model by showing an arrow pointing to θ_i

Box 2 Model Assumptions

For the analysis of batting abilities, we assume that a player's batting ability, θ_i, is constant for all at-bats, and that the outcome of any at-bat is independent of other at-bats. These assumptions may be false, but the notion of a constant underlying batting ability is a meaningful construct for our present purposes. Assumptions must be made for any statistical analysis, whether Bayesian or not, and the conclusions from any statistical analysis are conditional on its assumptions. An advantage of Bayesian analysis is that, relative to 20th century frequentist techniques, there is greater flexibility to make assumptions that are appropriate to the situation. For example, if we wanted to build a more elaborate analysis, we could incorporate data about when in the season the at-bats occurred, and estimate temporal trends in ability due to practice or fatigue. Or, we could incorporate data about which pitcher was being faced in each at-bat, and we could estimate pitcher difficulties simultaneously with batter abilities. But these elaborations, although possible in the Bayesian framework, would go far beyond our purposes in this chapter.

Kruschke, J. K. and Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. In: J. R. Busemeyer, Z. Wang, J. T. Townsend, and A. Eidels (Eds.), The Oxford Handbook of Computational and Mathematical Psychology, pp. 279-299. Oxford, UK: Oxford University Press.

from a beta distribution. The arrow is labeled with "$\sim \ldots i$" to indicate that the θ_i have credibilities distributed as a beta distribution for each of the individuals. The diagram shows beta distributions as they are conventionally parameterized by two shape parameters, denoted a_{pp} and b_{pp}, that can be algebraically redescribed in terms of the mean μ_{pp} and precision κ_{pp} of the distribution: $a_{pp} = \mu_{pp}\kappa_{pp}$ and $b_{pp} = (1 - \mu_{pp})\kappa_{pp}$.

To formally express our prior knowledge that all players, from all positions, are professionals in major league baseball, and, therefore, should mutually inform each other's estimates, we assume that the nine position abilities μ_{pp} come from an overarching beta distribution with mean $\mu_{\mu_{pp}}$ and precision $\kappa_{\mu_{pp}}$. This structure is illustrated in the upper part of Figure 13.1 by the split arrow, labeled with "$\sim \ldots pp$", pointing to μ_{pp} from a beta distribution. The value of $\mu_{\mu_{pp}}$ in the overarching distribution represents our estimate of the batting ability of major league players generally, and the value of $\kappa_{\mu_{pp}}$ represents how tightly clustered the abilities are across the nine positions. These across-position parameters are estimated from the data, along with all the other parameters.

The precisions of the nine distributions are also estimated from the data. The precisions of the position distributions, κ_{pp}, are assumed to come from an overarching gamma distribution, as illustrated in Figure 13.1 by the split arrow, labeled with "$\sim \ldots pp$", pointing to κ_{pp} from a gamma distribution. A gamma distribution is a generic and natural distribution for describing nonnegative values such as precisions (e.g., Kruschke, 2015). A gamma distribution is conventionally parameterized by shape and rate values, denoted in Figure 13.1 as $s_{\kappa_{pp}}$ and $r_{\kappa_{pp}}$. We assume that the precisions of each position can mutually inform each other; that is, if the batting abilities of catchers are tightly clustered, then the batting abilities or shortstops should probably also be tightly clustered, and so forth. Therefore the shape and rate parameters of the gamma distribution are themselves estimated.

At the top level in Figure 13.1 we incorporate any prior knowledge we might have about general properties of batting abilities for players in the

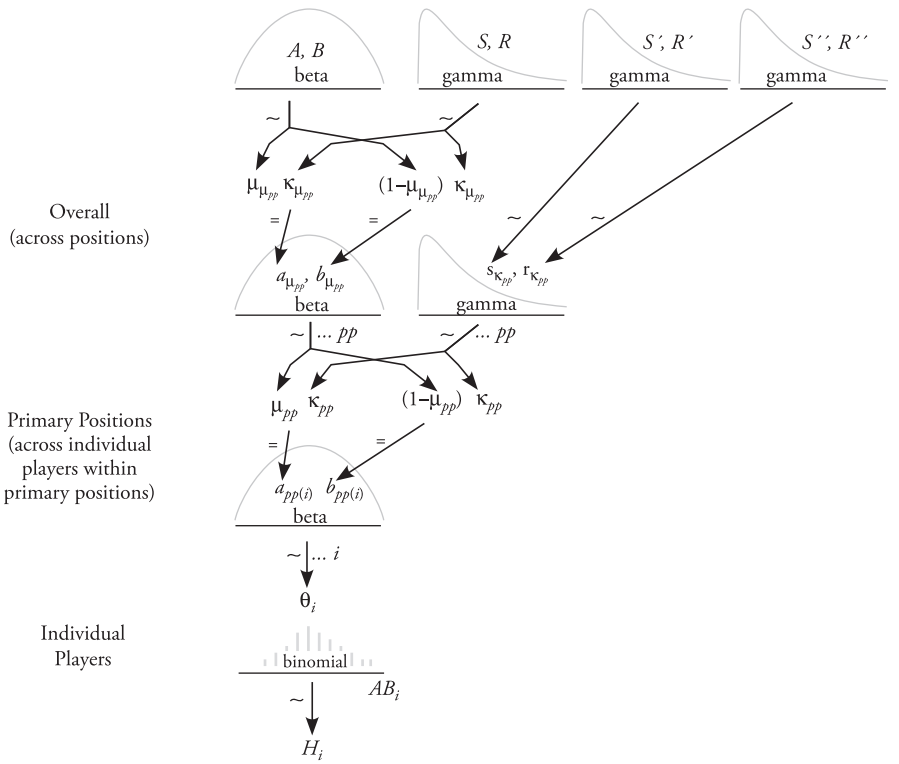

Fig. 13.1 The hierarchical descriptive model for baseball batting ability. The diagram should be scanned from the bottom up. At the bottom, the number of hits by the i^{th} player, H_i, are assumed to come from a binomial distribution with maximum value being the at-bats, AB_i, and probability of getting a hit being θ_i. See text for further details.

Kruschke, J. K. and Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. In: J. R. Busemeyer, Z. Wang, J. T. Townsend, and A. Eidels (Eds.), The Oxford Handbook of Computational and Mathematical Psychology, pp. 279-299. Oxford, UK: Oxford University Press.

major leagues, such as evidence from previous seasons of play. Baseball aficionados may have extensive prior knowledge that could be usefully implemented in a Bayesian model. Unlike baseball experts, we have no additional background knowledge, and, therefore, we will use very vague and noncommittal top-level prior distributions. Thus, the top-level beta distribution on the overall batting ability is given parameter values $A = 1$ and $B = 1$, which make it uniform over all possible batting abilities from zero to one. The top-level gamma distributions (on precision, shape, and rate) are given parameter values that make them extremely broad and noncommittal such that the data dominate the estimates, with minimal influence from the top-level prior.

There are 970 parameters in the model altogether: 948 individual θ_i, plus μ_{pp}, κ_{pp} for each of nine primary positions, plus μ_μ, κ_μ across positions, plus s_κ and r_κ. The Bayesian analysis yields credible *combinations* of the parameters in the 970-dimensional joint parameter space.

We care about the parameter values because they are meaningful. Our primary interest is in the estimates of individual batting abilities, θ_i, and in the position-specific batting abilities, μ_{pp}. We are also able to examine the relative precisions of abilities across positions to address questions such as, Are batting abilities of catchers as variable as batting abilities of shortstops? We will not do so here, however.

Results: Interpreting the Posterior Distribution

We used MCMC chains with total saved length of 15,000 after adaptation of 1,000 steps and burn-in of 1,000 steps, using 3 parallel chains called from the runjags package (Denwood, 2013), thinned by 30 merely to keep a modest file size for the saved chain. The diagnostics (see Box 1) assured us that the chains were adequate to provide an accurate and high-resolution representation of the posterior distribution. The effective sample size (ESS) for all the reported parameters and differences exceeded 6,000, with nearly all exceeding 10,000.

check of robustness against changes in top-level prior constants

Because we wanted the top-level prior distribution to be noncommittal and have minimal influence on the posterior distribution, we checked whether the choice of prior had any notable effect on the posterior. We conducted the analysis with different constants in the top-level gamma distributions, to check whether they had any notable influence on the resulting posterior distribution. Whether all gamma distributions used shape and rate constants of 0.1 and 0.1, or 0.001 and 0.001, the results were essentially identical. The results reported here are for gamma constants of 0.001 and 0.001.

comparisons of positions

We first consider the estimates of hitting ability for different positions. Figure 13.2, left side, shows the marginal posterior distributions for the μ_{pp} parameters for the positions of catcher and pitcher. The distributions show the credible values of the parameters generated by the MCMC chain. These marginal distributions collapse across all other parameters in the high-dimensional joint parameter space. The lower-left panel in Figure 13.2 shows the distribution of *differences* between catchers and pitchers. At every step in the MCMC chain, the difference between the credible values of μ_{catcher} and μ_{pitcher} was computed, to produce a credible value for the difference. The result is 15,000 credible differences (one for each step in the MCMC chain).

For each marginal posterior distribution, we provide two summaries: Its approximate mode, displayed on top, and its 95% *highest density interval* (HDI), shown as a black horizontal bar. A parameter value inside the HDI has higher probability density (i.e., higher credibility) than a parameter value outside the HDI. The total probability of parameter values within the 95% HDI is 95%. The 95% HDI indicates the 95% most credible parameter values.

The posterior distribution can be used to make discrete decisions about specific parameter values (as explained in Box 3). For comparing catchers and pitchers, the distribution of credible differences falls far from zero, so we can say with high credibility that catchers hit better than pitchers. (The difference is so big that it excludes any reasonable ROPE around zero that would be used in the decision rule described in Box 3.)

The right side of Figure 13.2 shows the marginal posterior distributions of the μ_{pp} parameters for the positions of right fielder and catcher. The lower-right panel shows the distribution of differences between right fielders and catchers. The 95% HDI of differences excludes a difference of zero, with 99.8% of the distribution falling above zero. Whether or not we reject zero as a credible

Kruschke, J. K. and Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. In: J. R. Busemeyer, Z. Wang, J. T. Townsend, and A. Eidels (Eds.), The Oxford Handbook of Computational and Mathematical Psychology, pp. 279-299. Oxford, UK: Oxford University Press.

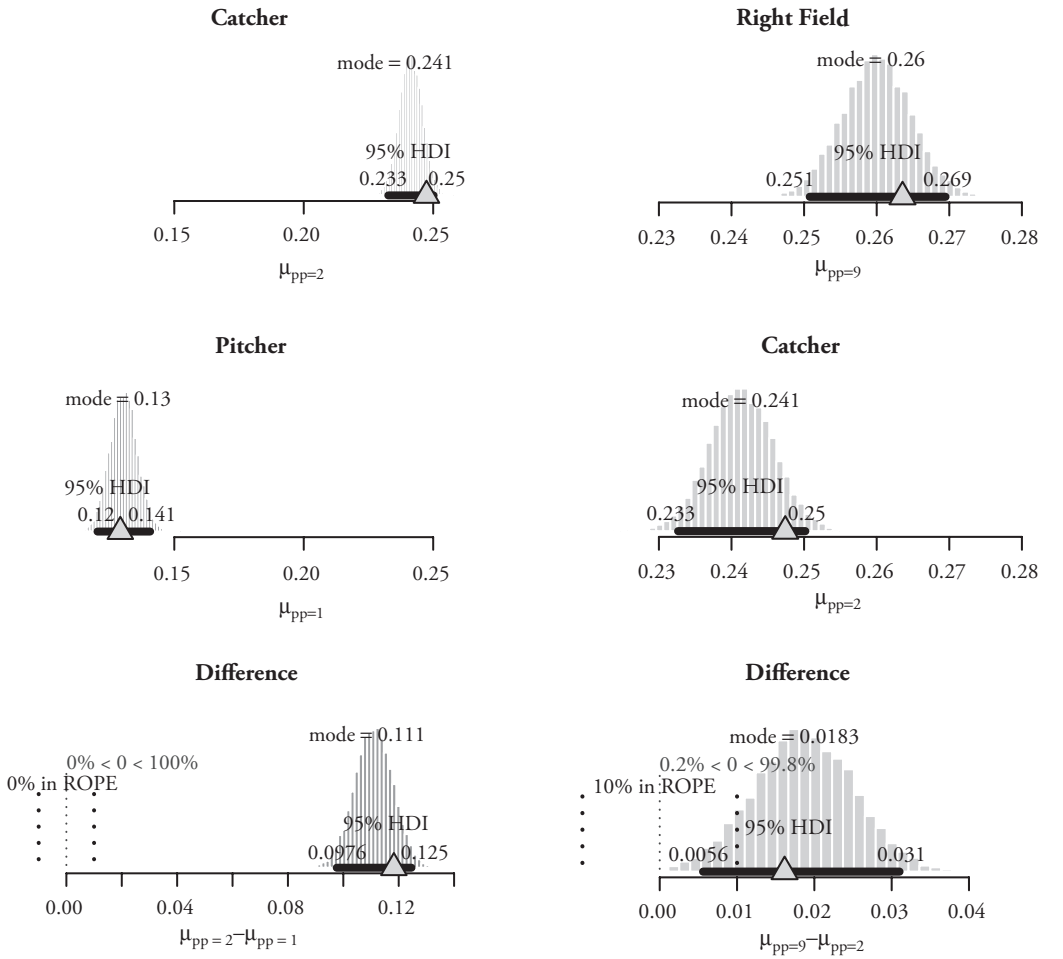

Fig. 13.2 Comparison of estimated batting abilities of different positions. In the data, there were 324 pitchers with a median of 4.0 at-bats, 103 catchers with a median of 170.0 at-bats, and 60 right fielders with a median of 340.5 at-bats, along with 461 players in six other positions. The modes and HDI limits are all indicated to three significant digits, with a trailing zero truncated from the display. In the lowest row, a difference of 0 is marked by a vertical dotted line annotated with the amount of the posterior distribution that falls below or above 0. The limits of the ROPE are marked with vertical dotted lines and annotated with the amount of the posterior distribution that falls inside it. The subscripts such as "$pp=2$" indicate arbitrary indexical values for the primary positions, such as 1 for pitcher, 2 for catcher, and so forth.

difference depends on our decision rule. If we use a ROPE from -0.01 to $+0.01$, as shown in Figure 13.2, then we would not reject a difference of zero because the 95% HDI overlaps the ROPE. The choice of ROPE depends on what is practically equivalent to zero as judged by aficionados of baseball. Our choice of ROPE shown here is merely for illustration.

In Figure 13.2, the triangle on the *x*-axis indicates the ratio in the data of total hits divided by total at-bats for all players in that position. Notice that the modes of the posterior are not centered exactly on the triangles. Instead, the modal estimates are *shrunken* toward the middle between the pitchers (who tend to have the lowest batting averages) and the catchers (who tend to have higher batting averages). Thus, the modes of the posterior marginal distributions are not as extreme as the proportions in the data (marked by the triangles). This shrinkage is produced by the mutual influence of data from all the other players, because they influence the higher-level distributions, which in turn influence the lower-level estimates. For example, the modal estimate for catchers is 0.241, which is less than the ratio of total hits to total at-bats for catchers. This shrinkage in the estimate for catchers is caused by the fact that there are 324 pitchers who, as a group, have relatively low batting

Kruschke, J. K. and Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. In: J. R. Busemeyer, Z. Wang, J. T. Townsend, and A. Eidels (Eds.), The Oxford Handbook of Computational and Mathematical Psychology, pp. 279-299. Oxford, UK: Oxford University Press.

Box 3 Decision Rules for Bayesian Posterior Distribution

The posterior distribution can be used for making decisions about the viability of specific parameter values. In particular, people might be interested in a landmark value of a parameter, or a difference of parameters. For example, we might want to know whether a particular position's batting ability exceeds 0.20, say. Or we might want to know whether two positions' batting abilities have a non-zero difference.

The decision rule involves using a *region of practical equivalence* (ROPE) around the null or landmark value. Values within the ROPE are equivalent to the landmark value for practical purposes. For example, we might declare that for batting abilities, a difference less than 0.04 is practically equivalent to zero. To decide that two positions have credibly different batting abilities, we check that the 95% HDI excludes the entire ROPE around zero. Using a ROPE also allows *accepting* a difference of zero: If the entire 95% HDI falls within the ROPE, it means that all the most credible values are practically equivalent to zero (i.e., the null value), and we decide to accept the null value for practical purposes. If the 95% HDI overlaps the ROPE, we withhold decision. Note that it is only the landmark value that is being rejected or accepted, not all the values inside the ROPE. Furthermore, the estimate of the parameter value is given by the posterior distribution, whereas the decision rule merely declares whether the parameter value is practically equivalent to the landmark value. We will illustrate use of the decision rule in the results from the actual analyses. In some cases we will not explicitly specify a ROPE, leaving some nonzero width ROPE implicit. In general, this allows flexibility in decision-making when limits of practical equivalence may change as competing theories and instrumentation change (Serlin & Lapsley. 1993). In some cases, the posterior distribution falls so far away from any reasonable ROPE that it is superfluous to specify a specific ROPE. For more information about the application of a ROPE, under somewhat different terms of "range of equivalence," "indifference zone," and "good-enough belt," see e.g., Carlin and Louis (2009); Freedman, Lowe, and Macaskill (1984); Hobbs and Carlin (2008); Serlin and Lapsley (1985, 1993); Spiegelhalter, Freedman, and Parmar (1994).

Notice that the decision rule is distinct from the Bayesian estimation itself, which produces the complete posterior distribution. We are using a decision rule only in case we demand a discrete decision from the continuous posterior distribution. There is another Bayesian approach to making decisions about null values that is based on comparing a "spike" prior on the landmark value against a diffuse prior, which we discuss in the final section on model comparison, but for the purposes of this chapter we focus on using the HDI with ROPE.

ability, and pull down the overarching estimate of batting ability for major-league players (even with the other seven positions taken into account). The overarching estimate in turn affects the estimate of all positions, and, in particular, pulls down the estimate of batting ability for catchers. We see in the upper right of Figure 13.2 that the estimate of batting ability for right fielders is also shrunken, but not as much as for catchers. This is because the right fielders tend to be at bat much more often than the catchers, and, therefore, the estimate of ability for right fielders more closely matches their data proportions. In the next section we examine results for individual players, and the concepts of shrinkage will become more dramatic and more clear.

comparisons of individual players

In this section we consider estimates of the batting abilities of individual players. The left side of Figure 13.3 shows a comparison of two individual players with the same record, 1 hit in 3 at-bats, but who play different positions, namely catcher and pitcher. Notice that the triangles are at the same place on the x-axes for the two players, but there are radically different estimates of their probability of getting a hit because of the different positions they play. The data from all the other catchers inform the model that catchers tend to have values of θ around 0.241. Because this particular catcher has so few data to inform his estimate, the estimate from the higher-level distribution dominates. The same is true for the pitcher, but the higher-level distribution says that pitchers tend to have values of θ around 0.130. The resulting distribution of differences, in the lowest panel, suggests that these two players have

Kruschke, J. K. and Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. In: J. R. Busemeyer, Z. Wang, J. T. Townsend, and A. Eidels (Eds.), The Oxford Handbook of Computational and Mathematical Psychology, pp. 279-299. Oxford, UK: Oxford University Press.

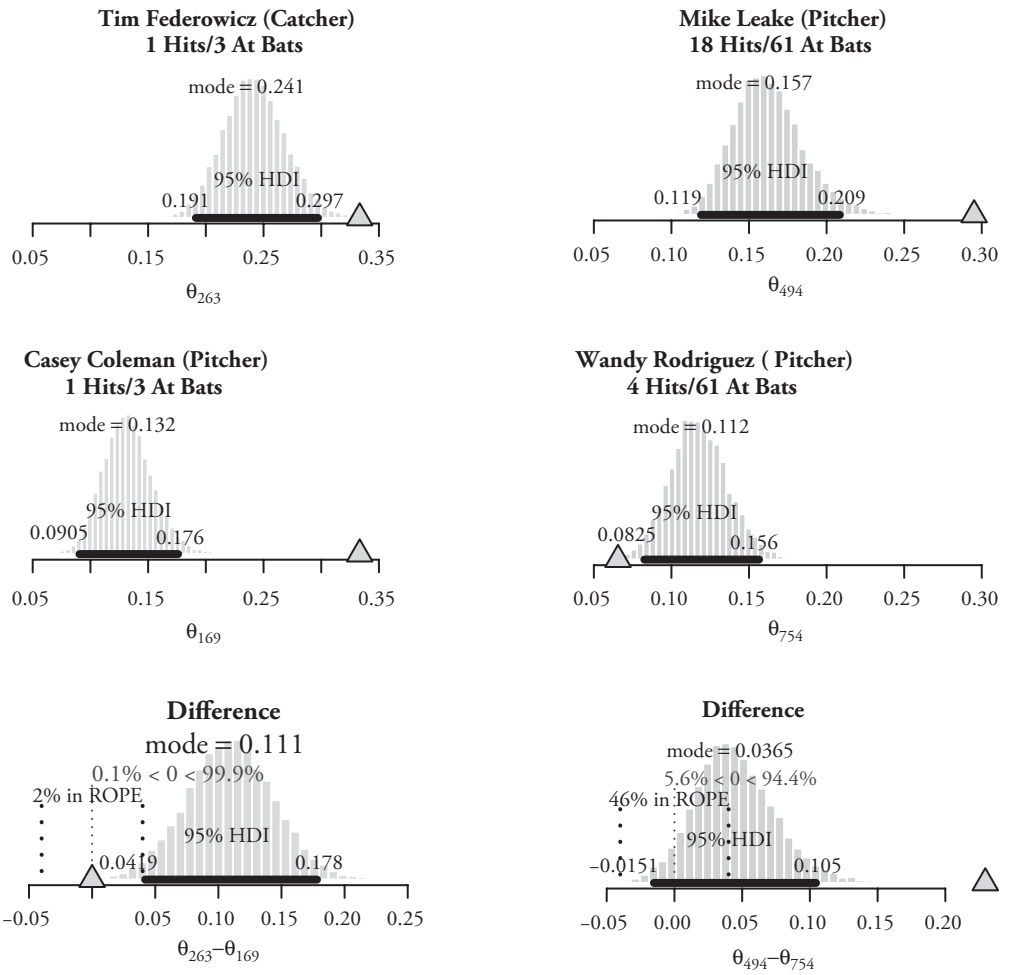

Fig. 13.3 Comparison of estimated batting abilities of different individual players. The left column shows two players with the same actual records of 1 hit in 3 at-bats, but very different estimates of batting ability because they play different positions. The right column shows two players with rather different actual records (18/61 and 4/61) but similar estimates of batting ability because they play the same position. Triangles show actual ratios of hits/at-bats. Bottom histograms display an arbitrary ROPE from −0.04 to +0.04; different decision makers might use a different ROPE. The subscripts on θ indicate arbitrary identification numbers of different players, such as 263 for Tim Federowicz.

credibly different hitting abilities, even though their actual hits and at-bats are identical. In other words, because we know the players play these particular different positions, we can infer that they probably have different hitting abilities.

The right side of Figure 13.3 shows another comparison of two individual players, both of whom are pitchers, with seemingly quite different batting averages of 18/61 and 4/61, as marked by the triangles on the *x*-axis. Despite the players' different hitting records, the posterior estimates of their hitting probabilities are not very different. Notice the dramatic shrinkage of the estimates toward the mode of players who are pitchers. Indeed, in the lower panel, we see that a difference of zero is credible, as it falls within the 95% HDI of the differences. The shrinkage is produced because there is a huge amount of data, from 324 pitchers, informing the position-level distribution about the hitting ability of pitchers. Therefore, the estimates of two individual pitchers with only modest numbers of at-bats are strongly shrunken toward the group-level mode. In other words, because we know that the players are both pitchers, we can infer that they probably have similar hitting abilities.

The amount of shrinkage depends on the amount of data. This is illustrated in Figure 13.4,

Kruschke, J. K. and Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. In: J. R. Busemeyer, Z. Wang, J. T. Townsend, and A. Eidels (Eds.), The Oxford Handbook of Computational and Mathematical Psychology, pp. 279-299. Oxford, UK: Oxford University Press.

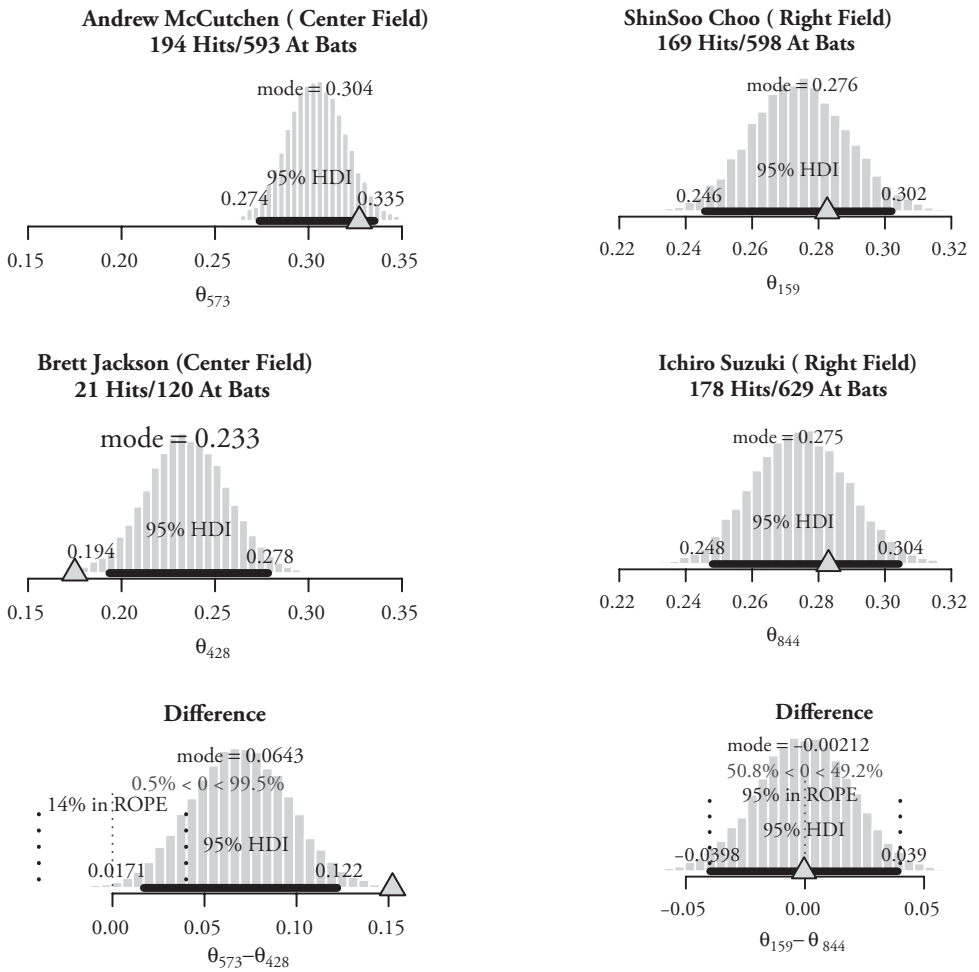

Fig. 13.4 The left column shows two individuals with rather different actual batting ratios (194/593 and 21/120) who both play center field. Although there is notable shrinkage produced by playing the same position, the quantity of data is sufficient to exclude a difference of zero from the 95% HDI on the difference (lower histogram); although the HDI overlaps the arbitrary ROPE shown here, different decision makers might use a different ROPE. The right column shows two right fielders with very high and nearly identical actual batting ratios. The 95% HDI of their difference falls within the ROPE in the lower right histogram. Note: Triangles show actual batting ratio of hits/at-bats.

which shows comparisons of players from the same position, but for whom there are much more personal data from more at-bats. In these cases, although there is some shrinkage caused by position-level information, the amount of shrinkage is not as strong because the additional individual data keep the estimates anchored closer to the data.

The left side of Figure 13.4 shows a comparison of two center fielders with 593 and 120 at-bats, respectively. Notice that the shrinkage of estimate for the player with 593 at-bats is not as extreme as the player with 120 at-bats. Notice also that the width of the 95% HDI for the player with 593 at-bats is narrower than for the player with 120 at-bats. This again illustrates the concept that the estimate is informed by both the data from the individual player and by the data from all the other players, especially those who play the same position. The lower left panel of Figure 13.4 shows that the estimated difference excludes zero (but still overlaps the particular ROPE used here).

The right side of Figure 13.4 shows right fielders with huge numbers of at-bats and nearly the same batting average. The 95% HDI of the difference falls almost entirely within the ROPE, so we might decide to declare that players have identical probability of getting a hit for practical purposes, that is, we might decide to accept the null value of zero difference.

Kruschke, J. K. and Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. In: J. R. Busemeyer, Z. Wang, J. T. Townsend, and A. Eidels (Eds.), The Oxford Handbook of Computational and Mathematical Psychology, pp. 279-299. Oxford, UK: Oxford University Press.

Shrinkage and Multiple Comparisons

In hierarchical models with multiple levels, there is shrinkage of estimates within each level. In the model of this section (Figure 13.1), there was shrinkage of the player-position parameters toward the overall central tendency, as illustrated by the pitcher and catcher distributions in Figure 13.2, and there was shrinkage of the individual-player parameters within each position toward the position central tendency, as shown by various examples in Figures 13.3 and Figure 13.4. The model also provided some strong inferences about player abilities based on position alone, as illustrated by the estimates for individual players with few at bats in the left column of Figure 13.3.

There were no corrections for multiple comparisons. We conducted all the comparisons without computing *p* values, and without worrying whether we might intend to make additional comparisons in the future, which is quite likely given that there are 9 positions and 948 players in whom we might be interested.

It is important to be clear that Bayesian methods do not prevent false alarms. False alarms are caused by accidental conspiracies of rogue data that happen to be unrepresentative of the true population, and no analysis method can fully mitigate false conclusions from unrepresentative data. There are two main points to be made with regard to false alarms in multiple comparisons from a Bayesian perspective.

First, the Bayesian method produces a posterior distribution that is fixed, given the data. The posterior distribution does not depend on which comparisons are intended by the analyst, unlike traditional frequentist methods. Our decision rule, using the HDI and ROPE, is based on the posterior distribution, not on a false alarm rate inferred from a null hypothesis and an intended sampling/testing procedure.

Second, false alarms are mitigated by shrinkage in hierarchical models (as exemplified in the right column of Figure 13.3). Because of shrinkage, it takes more data to produce a credible difference between parameter values. Shrinkage is a rational, mathematical consequence of the hierarchical model structure (which expresses our prior knowledge of how parameters are related) and the actually observed data. Shrinkage is not related in any way to corrections for multiple comparisons, which do not depend on the observed data but do depend on the intended comparisons. Hierarchical modeling is possible with non-Bayesian estimation, but frequentist decisions are based on auxiliary sampling distributions instead of the posterior distribution.

Example: Clinical Individual Differences in Attention Allocation

Hierarchical Bayesian estimation can be applied straightforwardly to more elaborate models, such as information processing models typically used in cognitive science. Generally, such models formally describe the processes underlying behavior in tasks such as thinking, remembering, perceiving, deciding, learning and so on. Cognitive models are increasingly finding practical uses in a wide variety of areas outside cognitive science. One of the most promising uses of cognitive process models is the field of *cognitive psychometrics* (Batchelder, 1998; Riefer, Knapp, Batchelder, Bamber, & Manifold, 2002; Vanpaemel, 2009), where cognitive process models are used as psychometric measurement models. These models have become important tools for quantitative clinical cognitive science (see Neufeld chapter 16, this volume).

In our second example of hierarchical Bayesian estimation, we use data from a classification task and a corresponding cognitive model to assess young women's attention to other women's body size and facial affect, following the research of Treat, Nosofsky, McFall, & Palmeri, (2002). Rather than relying on self-reports, Viken et a1. (2002) collected performance data in a prototype-classification task involving photographs of women varying in body size and facial affect. Furthermore, rather than using generic statistical models for data analysis, the researchers applied a computational model of category learning designed to describe underlying psychological properties. The model, known as the *multiplicative prototype model* (MPM: Nosofsky, 1987; Reed, 1972), has parameters that describe how much perceptual attention is allocated to body size or facial affect. The modeling made it possible to assess how participants in the task allocated their attention.

To measure attention allocation, Viken et al. (2002) tapped into women's perceived similarities of photographs of other women. The women in the photographs varied in their facial expressions of affect (happy to sad) and in their body size (light to heavy). We focus here on a particular categorization task in which the observer had to classify a target photo as belonging with reference photo X or with reference photo Y. In one version

Kruschke, J. K. and Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. In: J. R. Busemeyer, Z. Wang, J. T. Townsend, and A. Eidels (Eds.), The Oxford Handbook of Computational and Mathematical Psychology, pp. 279-299. Oxford, UK: Oxford University Press.

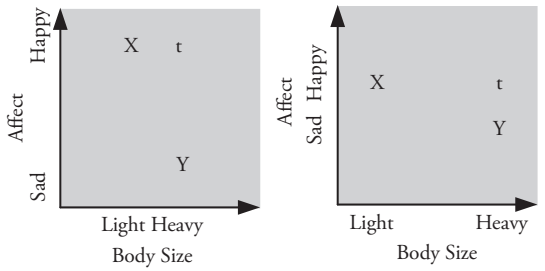

Fig. 13.5 The perceptual space for photographs of women who vary on body size (horizontal axis) and affect (vertical axis). Photo X shows a prototypical light, happy woman and photo Y shows a prototypical heavy, sad woman. The test photo, *t*, is categorized with X or Y according to its relative perceptual proximity to those prototypes. In the left panel, attention to body size (denoted *w* in the text) is low, resulting in compression of the body size axis, and, therefore, test photo *t* tends to be classified with prototype X. In the right panel, attention to body size is high, resulting in expansion of the body size axis, and, therefore, test photo *t* tends to be classified with prototype Y.

of the experiment, reference photo X was of a light, happy woman and reference photo Y was of a heavy, sad woman. In another version, not discussed here, the features of the reference photos were reversed. Suppose the target photo *t* showed a heavy, happy woman. If the observer was paying attention mostly to affect, then photo *t* should tend to be classified with reference photo X, which matched on affect. If the observer was paying attention mostly to body size, then photo *t* should tend to be classified with reference photo Y, which matched on body size. A schematic representation of the perceptual space for photographs is shown in Figure 13.5. In the actual experiment, there were many different target photos from throughout the perceptual space. By recording how each target photo was categorized by the observer, the observer's attention allocation can be inferred.

Viken et al. (2002) were interested in whether women suffering from the eating disorder, bulimia, allocated their attention differently than normal women. Bulimia is characterized by bouts of overconsumption of food with a feeling of loss of control, followed by self-induced vomiting or abuse of laxatives to prevent weight gain. The researchers were specifically interested in how bulimics allocated their attention to other women's facial affect and body size, because perception of body size has been the focus of past research into eating disorders, and facial affect is relevant to social perception but is not specifically implicated in eating disorders. An understanding of how bulimics allocate attention could have implications for both the etiology and treatment of the disease.

Viken et al. (2002) collected data from a group of woman who were high in bulimic symptoms, and from a group that was low. Viken et al. then used likelihood-ratio tests to compare a model that used separate attention weights in each group to a model that used a single attention weight for both groups. Their model-comparison approach revealed that high-symptom women, relative to low-symptom women, display enhanced attention to body size and decreased attention to facial affect.

In contrast to their non-Bayesian, nonhierarchical, nonestimation approach, we use a Bayesian hierarchical estimation approach to investigate the same issue. The hierarchical nature of our approach means that we do not assume that all subjects within a symptom group have the same attention to body size. Bayesian inference and decision-making implies that we do not require assumptions about sampling intentions and multiple tests that are required for computing *p* values. Moreover, our use of estimation instead of only model comparison ensures that we will know how much the groups differ.

The Data

Viken et a1. (2002) obtained classification judgments from 38 women on 22 pictures of other women, varying in body size (light to heavy) and facial affect (happy to sad). Symptoms of bulimia were also measured for all of the women. Eighteen of these women had BULIT scores exceeding 88, which is considered to be high in bulimic symptoms (Smith & Thelen, 1984). The remaining 20 women had BULIT scores lower than 45, which is considered to be low in bulimic symptoms. Each woman performed the classification task described earlier, in which she was instructed to freely classify each target photo *t* as one of two types of women exemplified by reference photo X and reference photo Y. No corrective feedback was provided. Each target photo was presented twice, hence, for each woman *i*, the data include the frequency of classifying stimulus *t* as a type X, ranging between 0 and 2. Our goal is to use these data to infer a meaningful measure of attention allocation for each individual observer, and simultaneously to infer an overall measure of attention allocation for women high in bulimic symptoms and for women low in bulimic symptoms. We will rely

Kruschke, J. K. and Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. In: J. R. Busemeyer, Z. Wang, J. T. Townsend, and A. Eidels (Eds.), The Oxford Handbook of Computational and Mathematical Psychology, pp. 279-299. Oxford, UK: Oxford University Press.

on a hierarchical extension of the MPM, as described next.

The Descriptive Model with Its Meaningful Parameters

Models of categorization take perceptual stimuli as input and generate precise probabilities of category assignments as output. The input stimuli must be represented formally, and many leading categorization models assume that stimuli can be represented as points in a multidimensional space, as was suggested in Figure 13.5. Importantly, the models assume that attention plays a key role in categorization, and formalize the attention allocated to perceptual dimensions as free parameters (for a review see, e.g., Kruschke, 2008). In particular, the MPM (Nosofsky, 1987) determines the similarity between a target item and a reference item by multiplicatively weighting the separation of the items on each dimension by the corresponding attention allocated to each dimension. The higher the similarity of a stimulus to a reference category prototype, relative to other category prototypes, the higher the probability of assigning the stimulus to the reference category.

For each trial in which a target photo t is presented with reference photos X and Y, the MPM produces the probability, $p_i(X|t)$, that the i^{th} observer classifies stimulus t as category X. This probability depends on two free parameters. One parameter is denoted w_i, which indicates the attention that the i^{th} observer pays to body size. The value of w_i can range from 0 to 1. Attention to affect is simply $1 - w_i$. The second parameter is denoted c_i and called the "sensitivity" of observer i. The sensitivity can be thought of as the observer's decisiveness, which is how strongly the observer converts a small similarity advantage for X into a large choice advantage for X. Note that attention and sensitivity parameters can differ across observers, but not across stimuli, which are assumed to have fixed locations in an underlying perceptual space.

Formally, the MPM posits that the probability that photo t will be classified with reference photo X instead of reference photo Y is determined by the similarity of t to X relative to the total similarity:

$$p_i(X|t) = s_{tX}/(s_{tX} + s_{tY}). \tag{3}$$

The similarity between target and reference is, in turn, determined as a nonlinearly decreasing function of distance between t and X, d_{tX}, in the psychological space:

$$s_{tX} = \exp(-c_i\, d_{tX}) \tag{4}$$

where $c_i > 0$ is the sensitivity parameter for observer i. The psychological distance between target t and reference X is given by the weighted distance between the corresponding points in the 2-dimensional psychological space:

$$d_{tX} = \left[w_i \left| x_{tb} - x_{Xb} \right|^2 + (1 - w_i) \left| x_{ta} - x_{Xa} \right|^2 \right]^{1/2}, \tag{5}$$

where x_{ta} denotes the position of the target on the affect dimension, and x_{tb} denotes the position of the target on the body-size dimension. These positions are normative average ratings of the photographs on two 10-point scales: body size (1 = underweight, 10 = overweight), and affect (1 = unhappy, 10 = happy), as provided by a separate sample of young women. The free parameter $0 < w_i < 1$ corresponds to the attention weight on the body size dimension for observer i. It reflects the key assumption of the MPM that the structure of the psychological space is systematically modified by selective attention (see Figure 13.5).

HIERARCHICAL STRUCTURE

We construct a hierarchical model that has parameters to describe each individual, and parameters to describe the overall tendencies of the bulimic and normal groups. The hierarchy is analogous to the baseball example discussed earlier: Just as individual players were nested within fielding positions, here individual observers are nested within bulimic symptom groups. (One difference, however, is that we do not build an overarching distribution across bulimic-symptom groups because there are only two groups.) With this hierarchy, we express our prior expectation that bulimic women are similar but not identical to each other, and nonbulimic women are similar but not identical to each other, but the two groups may be different.

The hierarchical model allows the parameter estimates for an individual observer to be rationally influenced by the data from other individuals within their symptom group. In our model, the individual attention weights are assumed to come from an overarching distribution that is characterized by a measure of central tendency and of dispersion. The overarching distributions for the high-symptom and low-symptom groups are estimated separately. As the attention weights w_i are constrained to range between 0 and 1,

Kruschke, J. K. and Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. In: J. R. Busemeyer, Z. Wang, J. T. Townsend, and A. Eidels (Eds.), The Oxford Handbook of Computational and Mathematical Psychology, pp. 279-299. Oxford, UK: Oxford University Press.

we assume the parent distribution for the w_i's is a beta distribution, parameterized by mean $\mu_w^{[g]}$ and precision $\kappa_w^{[g]}$, where $[g]$ indexes the group membership (i.e., high symptom or low symptom). The individual sensitivities, c_i, are also assumed to come from an overarching distribution. Since the sensitivities are non-negative, a gamma distribution is a convenient parent distribution, parameterized by mode $mo_c^{[g]}$ and standard deviation $\sigma_c^{[g]}$, where $[g]$ again indicates the group membership (i.e., high symptom or low symptom). The group-level parameters (i.e., $\mu_w^{[g]}$, $mo_c^{[g]}$, $\kappa_w^{[g]}$ and $\sigma_c^{[g]}$) are assumed to come from vague, noncommittal uniform distributions. There are 84 parameters altogether, including w_i and c_i for 38 observers and the 8 group level parameters. Figure 13.6 summarizes the hierarchical model in an integrated diagram. The caption provides details.

The parameters of most interest are the group-level attention to body size, $\mu_w^{[g]}$, for $g \in \{\text{low}, \text{high}\}$. Other meaningful questions could focus on the relative variability among groups in attention, which would be addressed by considering the $\kappa_w^{[g]}$ parameters, but we will not pursue these here.

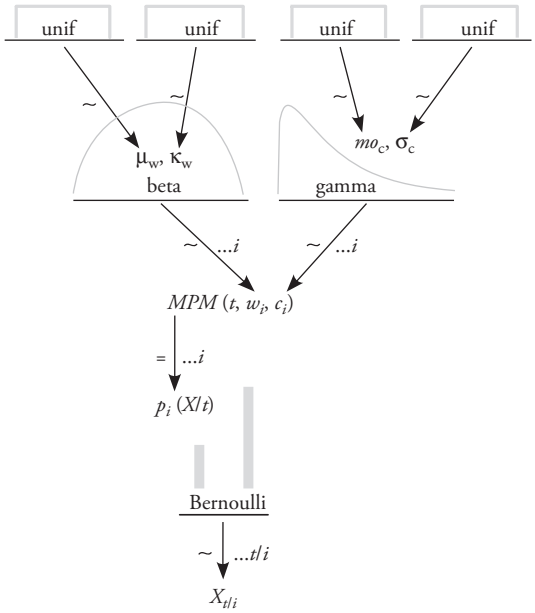

Fig. 13.6 The hierarchical model for attention allocation. At the bottom of the diagram, the classification data are denoted as $X_{t|i} = 1$ if observer i says "X" to target t, and $X_{t|i} = 0$ otherwise. The responses come from a Bernoulli distribution that has its success parameter determined by the MPM, as defined in Eqs. 3, 4, and 5 in the main text. The ellipsis on the arrow pointing to the response indicates that this relation holds for all targets within every individual. Scanning up the diagram, the individual attention parameters, w_i, come from an overarching group-level beta distribution that has mean μ_w and concentration κ_w (hence shape parameters of $a_w = \mu_w\kappa_w$ and $b_w = (1 - \mu_w)\kappa_w$, as was indicated explicitly for the beta distributions in Figure 13.1). The individual sensitivity parameters c_i come from an overarching group-level gamma distribution that has mode mo_c and standard deviation σ_c (with shape and rate parameters that are algebraic combinations of mo_c and σ_c; see Kruschke, 2015, Section 9.2.2). The group-level parameters all come from noncommittal, broad uniform distributions. This model is applied separately to the high-symptom and low-symptom observers.

Results: Interpreting the Posterior Distribution

The Bayesian hierarchical approach to estimation yields attention weights for each observer, informed by all the other observers in the group. At the same time, it provides an estimate of the attention weight at the group level. Further, for every individual estimate and the group level estimates, a measure of uncertainty is provided, in the form of a credible interval (95% HDI), which can be used as part of a decision rule to decide whether or not there are credible differences between individuals or between groups.

The MCMC process used 3 chains with a total of 100,000 steps after a burn-in of 4,000 steps. It produced a smooth (converged) representation of the 84-dimensional posterior distribution. We use the MCMC sample as an accurate and high-resolution representation of the posterior distribution.

CHECK OF ROBUSTNESS AGAINST CHANGES IN TOP-LEVEL PRIOR CONSTANTS

We conducted a sensitivity analysis by using different constants in the top-level uniform distributions, to check whether they had any notable influence on the resulting posterior distribution. Whether all uniform distributions assumed an upper bound of 10 or 50, the results were essentially identical. The results reported here are for an upper bound of 10.

COMPARISON ACROSS GROUPS OF ATTENTION TO BODY SIZE

Figure 13.7 shows the marginal posterior distribution for the group-level parameters of most interest. The left side shows the distribution of the central tendency of attention to body size for each group as well as the distribution of their difference. In particular, the bottom left histogram shows that the low-symptom group has an attention weight on body size about 0.36 lower than the high-symptom

Kruschke, J. K. and Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. In: J. R. Busemeyer, Z. Wang, J. T. Townsend, and A. Eidels (Eds.), The Oxford Handbook of Computational and Mathematical Psychology, pp. 279-299. Oxford, UK: Oxford University Press.

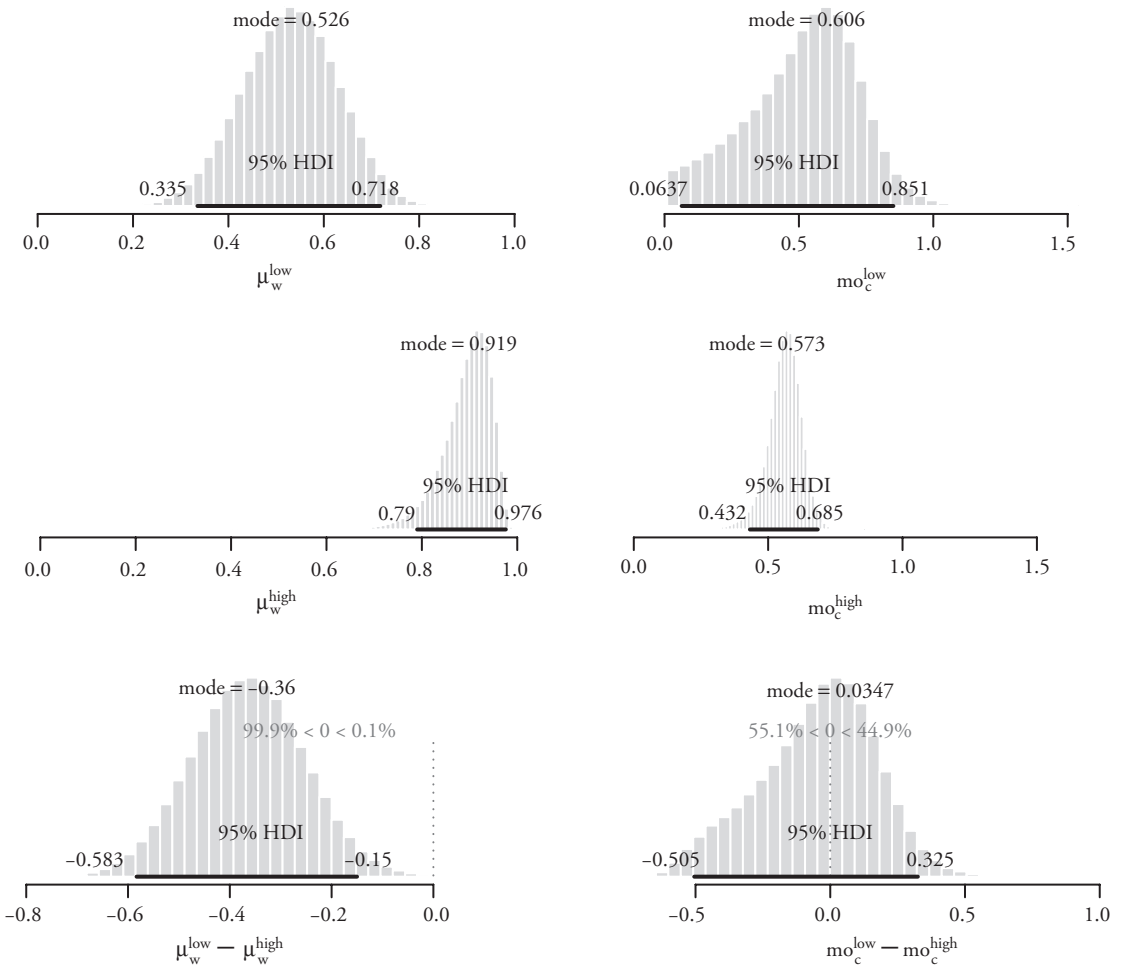

Fig. 13.7 Marginal posterior distribution of group-level parameters for the prototype classification task. The left column shows the group-level central tendency of the attention weight on body size, $\mu_w^{[g]}$. The bottom-left histogram reveals a credibly nonzero difference between groups, with low-symptom observers allocating about 0.36 less attention to body-size than high-symptom observers. The 95% HDI is so far away from a difference of zero that any reasonable ROPE would be excluded; therefore, we do not specify a particular ROPE. The right column shows the group-level central tendency of the sensitivity parameter, $mo_c^{[g]}$. The bottom-right histogram shows that zero difference is squarely among the most credible differences.

group, and this difference is credibly nonzero. The right side shows that the most credible difference of sensitivities is near zero.

The conclusions from our hierarchical Bayesian estimation agree with those of Viken et al. (2002), who took a non-Bayesian, nonhierarchical, model-comparison approach. We also find that high-symptom women, relative to low-symptom women, show enhanced attention to body size and decreased attention to facial affect, but no differences in their sensitivities. However, our hierarchical Bayesian estimation approach has provided explicit distributions on the credible differences between the groups.

comparisons across individual women's attention to body size

Although the primary question of interest involves the group-level central tendencies, hierarchical Bayesian estimation also automatically provides estimates of the attention weights of individual women. Figure 13.8 shows the estimates of individual attention weights w_i for three women, based on the hierarchical Bayesian estimation that shares information across all observers to inform the estimate of each individual observer. Figure 13.8 also shows the individual estimates from a nonhierarchical MLE, which derives each individual estimate from the data of a single observer only.

Kruschke, J. K. and Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. In: J. R. Busemeyer, Z. Wang, J. T. Townsend, and A. Eidels (Eds.), The Oxford Handbook of Computational and Mathematical Psychology, pp. 279-299. Oxford, UK: Oxford University Press.

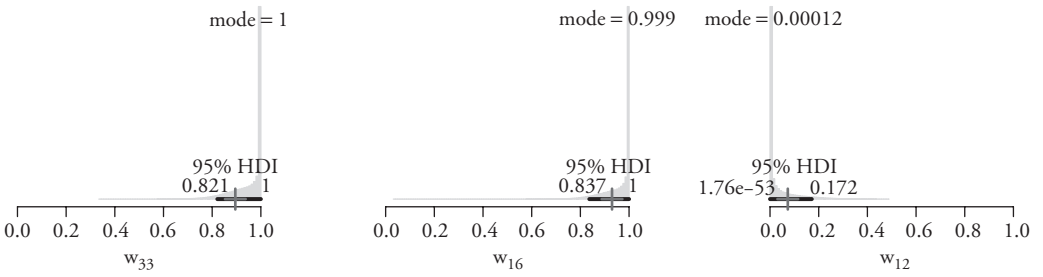

Fig. 13.8 Posterior of attention weights w_i of three individual observers. The vertical mark on the HDI indicates the MLE of the attention weight based on the individual's data only. Observer 33 is a high-symptom woman, whose estimate is shrunk upward (toward one). Observers 16 and 12 are both low-symptom women, whose estimates are shrunken in different directions (upwards for 16, downwards for 12).

Figure 13.8 illustrates that in hierarchical models, data from one individual influence inferences about the parameters of the other individuals. Technically, this happens because each woman's data influence the group-level parameters, which affect all the individual-level parameter estimates. For example, the hierarchical Bayesian modal estimate of the attention weight for observer 33, a high-symptom woman, is 1, which is larger than the nonhierarchical MLE of 0.89. This shrinkage in the hierarchical estimate is caused by the fact that most other high-symptom women tend to have relatively high attention weights, thereby pulling up the group-level estimate of the attention weight and the estimates for each individual high-symptom woman. Shrinkage also occurs for the low-symptom woman, shown in the other panels of Figure 13.8. The second panel shows that for observer 16, the hierarchical Bayesian modal estimate of the attention weight is 1, which is higher than the nonhierarchical estimate of 0.93. For observer 12, however, shrinkage is in the opposite direction: the hierarchical Bayesian modal estimate of the attention weight is smaller than the MLE based on individual data (0 vs 0.07). These opposite directions in shrinkage of the estimate are caused by the fact that the overarching beta distribution for low-symptom women is bimodal (i.e., have shape parameters less than 1.0), with one mode near 0 and a second mode near 1, indicating that low-symptom women tend to have either a low attention weight or a high attention weight. This bimodality is evident in the data and is not merely an artifact of the model, insofar as many women classify as if paying most attention to either one dimension or the other. Woman with MLE's close to 0 have hierarchical Bayesian estimates even closer to 0, whereas woman with MLE's close to 1 have hierarchical Bayesian estimates even closer to 1. Shrinkage for the low-symptom women thus highlights that shrinkage is not necessarily inward, toward the middle of the higher-level distribution; it can also be outward, and always toward the modes.

Model Comparison as a Case of Estimation in Hierarchical Models

In the examples discussed earlier, Bayesian estimation was the reallocation of credibility across the space of parameter values, for continuous parameters. We can think of each distinct parameter value (or joint combination of values in a multiparameter space) as a distinct model of the data. Because the parameter values are on a continuum, there is a continuum of models. Under this conceptualization, Bayesian parameter estimation is model comparison for an infinity of models.

Often, however, people may think of different models as being distinct, discrete descriptions, not on a continuum. This conceptualization of models makes little difference from a Bayesian perspective. When models are discrete, there is still a parameter that relates them to each other, namely an indexical parameter that has value 1 for the first model, 2 for the second model, and so on. Bayesian model comparison is then Bayesian estimation, as the reallocation of credibility across the values of the indexical parameter. The posterior probabilities of the models are simply the posterior probabilities of the indexical parameter values. Bayesian inference operates, mathematically, the same way regardless of whether the parameter that relates models is continuous or discrete.

Figure 13.9 shows a hierarchical diagram for comparing two models. At the bottom of the

Kruschke, J. K. and Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. In: J. R. Busemeyer, Z. Wang, J. T. Townsend, and A. Eidels (Eds.), The Oxford Handbook of Computational and Mathematical Psychology, pp. 279-299. Oxford, UK: Oxford University Press.

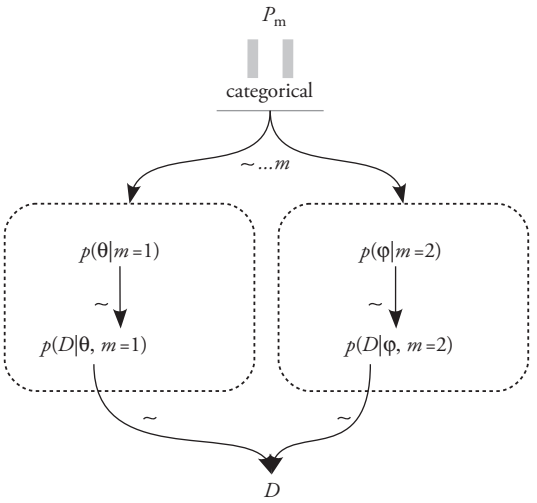

Fig. 13.9 Model comparison as hierarchical modeling. Each dashed box encloses a discrete model of the data, and the models depend on a higher-level indexical parameter at the top of the diagram. See text for further details.

diagram are the data, D. Scanning up the diagram, the data are distributed according to likelihood function $p(D|\theta, m = 1)$ when the model index m is 1. The likelihood function for model 1 involves a parameter θ, which has a prior distribution specified by $p(\theta|m=1)$. All the terms involving the parameter θ are enclosed in a dashed box, which indicates the part of the overall hierarchy that depends on the higher-level indexical parameter, m, having value $m = 1$. Notice, in particular, that the prior on the parameter θ is an essential component of the model; that is, the model is not only the likelihood function but also the prior. When $m = 2$, the data are distributed according to the model on the right of the diagram, involving a likelihood function and prior with parameter ϕ. At the top of the hierarchy is a categorical distribution that specifies the prior probability of each indexical value of m, that is, the prior probability of each model as a discrete entity. This hierarchical diagram is analogous to previous hierarchical diagrams in Figures 13.1 and Figure 13.6, but the top-level distribution is discrete and lower-level parameters and structures can change discretely instead of continuously when the top-level parameter value changes.

The sort of hierarchical structure diagrammed in Figure 13.9 can be implemented in the same MCMC sampling software we used for the baseball and categorization examples earlier. The MCMC algorithm generates representative values of the indexical parameter m, together with representative values of the parameter θ (when $m = 1$) and the parameter ϕ (when $m = 2$). The posterior probability of each model is approximated accurately by the proportion of steps that the MCMC chain visited each value of m. For a hands-on introduction to MCMC methods for Bayesian model comparison, see Chapter 10 of Kruschke (2015) and Lodewyckx et al. (2011). Examples of Bayesian model comparison are also provided by Vandekerckhove, Matzke, and Wagenmakers in chapter 14, this volume.

When comparing models, it is crucially important to set appropriately the prior distributions within each model, because the estimation of the model index can be very sensitive to the choice of prior. In the context of Figure 13.9, we mean that it is crucial to set the prior distributions, $p(\theta|m=1)$ and $p(\phi|m=2)$, so that they accurately express the priors intended for each model. Otherwise it is trivially easy to favor one model over the other, perhaps inadvertently, by setting one prior to values that accommodate the data well while setting the other prior to values that do not accommodate the data well. If each model comes with a theory or previous research that specifically informs the model, then that theory or research should be used to set the prior for the model. Otherwise, the use of generic default priors can unwittingly favor one model over the other. When there are not strong theories or previous research to set the priors for each model, a useful approach for setting priors is as follows: Start each model with vague default priors. Then, using some modest amount of data that represent consensually accepted previous findings, update all models with those data. The resulting posterior distributions in each model are then used as the priors for the model comparison, using the new data. The priors, by being modestly informed, have mitigated the arbitrary influence of inappropriate default priors, and have set the models on more equal playing fields by being informed by the same prior data. These and other issues are discussed in the context of cognitive model comparison by Vanpaemel (2010) and Vanpaemel and Lee (2012).

A specific case of the hierarchical structure in Figure 13.9 occurs when the two models have the same likelihood function, and hence the same parameters, but different prior distributions. In this case, the model comparison is really a comparison of two competing choices of prior distribution for the parameters. A common application for this specific

Kruschke, J. K. and Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. In: J. R. Busemeyer, Z. Wang, J. T. Townsend, and A. Eidels (Eds.), The Oxford Handbook of Computational and Mathematical Psychology, pp. 279-299. Oxford, UK: Oxford University Press.

case is null hypothesis testing. The null hypothesis is expressed as a prior distribution with all its mass at a single value of the parameter, namely the "null" value, such as $\theta = 0$. If drawn graphically, the prior distribution would look like a spike-shaped distribution. The alternative hypothesis is expressed as a prior distribution that spreads credibility over a broad range of the parameter space. If drawn graphically, the alternative prior might resemble a thin (i.e., short) slab-shaped distribution. Model comparison then amounts to the posterior probabilities of the spike-shaped (null) prior and the slab-shaped (alternative) prior. This approach to null-hypothesis assessment depends crucially on the meaningfulness of the chosen alternative-hypothesis prior, because the posterior probability of the null-hypothesis prior is not absolute but merely relative to the chosen alternative-hypothesis prior. The relative probability of the null-hypothesis prior can change dramatically for different choices of the alternative-hypothesis prior. Because of this sensitivity to the alternative-hypothesis prior, we recommend that this approach to null-hypothesis assessment is used only with caution when it is clearly meaningful to entertain the possibility that the null value could be true and a justifiable alternative-hypothesis prior is available. In such cases, the prior-comparison approach can be very useful. However, in the absence of such meaningful priors, null-value assessment most safely proceeds by explicit estimation of parameter values within a single model, with decisions about null values made according to the HDI and ROPE as exemplified earlier. For discussion of Bayesian approaches to null-value assessment, see, for example, Kruschke (2011), Kruschke (2013, Appendix D), Morey and Rouder (2011), Wagenmakers (2007), and Wetzels, Raaijmakers, Jakab, and Wagenmakers (2009).

Conclusion

In this chapter we discussed two examples of hierarchical Bayesian estimation. The baseball example (Figure 13.1) illustrated multiple levels with shrinkage of estimates within each level. We chose this example because it clearly illustrates the effects of hierarchical structure in rational inference of individual and group parameters. The categorization example (Figure 13.6) illustrated the use of hierarchical Bayesian estimation for psychometric assessment via a cognitive model. The parameters are meaningful in the context of the cognitive model, and Bayesian estimation provides a complete posterior distribution of credible parameter values for individuals and groups. Other examples of hierarchical Bayesian estimation can be found, for instance, in articles by Bartlema, Lee, Wetzels, and Vanpaemel (2014), Lee (2011), Rouder and Lu (2005), Rouder, Lu, Speckman, Sun, and Jiang (2005), and Shiffrin, Lee, Kim, and Wagcnmakers (2008).

The hierarchical Bayesian method is very attractive because it allows the analyst to define meaningfully structured models that are appropriate for the data. For example, there is no artificial dilemma of deciding between doing separate individual analyses or collapsing across all individuals, which both have serious shortcomings (Cohen, Sanborn, & Shiffrin: 2008). When collapsing the data across participants in each group, it is implicitly assumed that all participants within a group behave identically. Such an assumption is often untenable. The other extreme of analyzing every individual separately with no pooling across individuals can be highly error prone, especially when each participant contributed only small amounts of data. A hierarchical analysis provides a middle ground between these two strategies, by acknowledging that people are different, without ignoring the fact that they represent a common group or condition. The hierarchical structure allows information provided by one participant to flow rationally to the estimates of other participants. This sharing of information across participants via hierarchical structure occurs in both the classification and baseball examples of this chapter.

A second key attraction of hierarchical Bayesian estimation is that software for expressing complex, nonlinear hierarchical models (e.g., Lunn et al. 2000, Plummer 2003, Stan Development Team 2012), produces a complete posterior distribution for direct inferences about credible parameter values without need for p values or corrections for multiple comparisons. The combination of ease of defining specifically appropriate models and ease of direct inference from the posterior distribution makes hierarchical Bayesian estimation an extremely useful approach to modeling and data analysis.

Acknowledgments

The authors gratefully acknowledge Rick Viken and Teresa Treat for providing data from Viken, Treat, Nosofsky, McFall, and Palmeri (2002). Appreciation is also extended to E.-J.

Kruschke, J. K. and Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. In: J. R. Busemeyer, Z. Wang, J. T. Townsend, and A. Eidels (Eds.), The Oxford Handbook of Computational and Mathematical Psychology, pp. 279-299. Oxford, UK: Oxford University Press.

Wagenmakers and two anonymous reviewers who provided helpful comments that improved the presentation. Correspondence can be addressed to John K. Kruschke, Department of Psychological and Brain Sciences, Indiana University, 1101 E. 10th St., Bloomington IN 47405-7007, or via electronic mail to kruschke@indiana.edu. Supplementary information about Bayesian data analysis can be found at http://www.indiana.edu/~kruschke/

Notes

1. The most general definition of a confidence interval is the range of parameter values that would not be rejected according to a criterion p value, such as $p < 0.05$. These limits depend on the arbitrary settings of other parameters, and can be difficult to compute.
2. Data retrieved December 22, 2012 from http://www.baseball-reference.com/leagues/MLB/2012-standard-batting.shtml
3. This analysis was summarized at http://doingbayesiandataanalysis.blogspot.com/2012/11/shrinkage-in-multi-level-hierarchical.html
4. In the context of a normal distribution, instead of a beta distribution, the "precision" is the reciprocal of variance. Intuitively, it refers to the narrowness of the distribution for either the normal or beta distributions.

Glossary

Hierarchical model: A formal model that can be expressed such that one parameter is dependent on another parameter. Many models can be meaningfully factored this way, for example when there are parameters that describe data from individuals, and the individual-level parameters depend on group-level parameters.

Highest density interval (HDI): The highest density interval summarizes the interval under a probability distribution where the probability densities inside the interval are higher than probability densities outside the interval. A 95% HDI includes the 95% of the distribution with the highest probability density.

Markov chain Monte Carlo (MCMC): A class of stochastic algorithms for obtaining samples from a probability distribution. The algorithms take a random walk through parameter space, favoring values that have higher probability. With a sufficient number of steps, the values of the parameter are visited in proportion to their probabilities and therefore the samples can be used to approximate the distribution. Widely used examples of MCMC are the Gibbs sampler and the Metropolis-Hastings algorithm.

Posterior distribution: A probability distribution over parameters derived via Bayes' rule from the prior distribution by taking into account the targeted data.

Prior distribution: A probability distribution over parameters representing the beliefs, knowledge or assumptions about the parameters without reference to the targeted data. The prior distribution and the likelihood function together define a model.

Region of practical equivalence (ROPE): An interval around a parameter value that is considered to be equivalent to that value for practical purposes. The ROPE is used as part of a decision rule for accepting or rejecting particular parameter values.

References

Bartlema, A., Lee, M D., Wetzels, R., & Vanpaemel, W. (2014). A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. *Journal of Mathematical Psychology, 59,* 132–150.

Batchelder, W H. (1998). Multinomial processing tree models and psychological assessment. *Psychological Assessment, 10,* 331–344.

Bayes, T., & Price, R. (1763). An essay towards solving a problem in the doctrine of chances. By the Late Rev. Mr. Bayes, F.R.S. Communicated by Mr. Price, in a Letter to John Canton, A.M.F.R.S. *Philosophical Transactions, 53,* 370–418. doi: 10.1098/rstl.1763.0053

Carlin, B P., & Louis, T. A. (2009). Bayesian methods for data analysis (3rd ed.). Boca Raton, FL: CRC Press.

Cohen, A. L., Sanborn, A. N., & Shiffrin, R, M. (2008). Model evaluation using grouped or individual data. *Psychonomic Bulletin & Review, 15,* 692–712.

Denwood, M. J. (2013). runjags: An R package providing interface utilities, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*, (in review). http://cran.r-project.org/web/bpackages/runjags/

Doyle, A. C. (1890). The sign of four. London, England: Spencer Blackett.

Freedman, L. S., Lowe, D., & Macaskill, P. (1984). Stopping rules for clinical trials incorporating clinical opinion. *Biometrics, 40,* 575–586.

Hobbs, B. P., & Carlin, B. P. (2008). Practical Bayesian design and analysis for drug and device clinical trials. *Journal of Biopharmaceutical Statistics, 18*(1), 54–80.

Kruschke, J. K. (2008). Models of categorization. R. Sun (Ed.), *The Cambridge Handbook of Computational Psychology* (p. 267–301). New York, NY: Cambridge University Press.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6*(3) 299–312.

Kruschke, J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General, 142*(2), 573–603. doi: 10.1037/a0029146

Kruschke, J. K. (2015). Doing Bayesian data analysis, Second edition: A tutorial with R, JAGS, and Stan. Waltham, Academic Press/Elsevier.

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology, 55,* 1–7.

Lodewyckx, T., Kim, W., Lee, M. D., Tuerlinckx, F., Kuppens, P., & Wagenmakers, E. J. (2011). A tutorial on Bayes factor estimation with the product space method. *Journal of Mathematical Psychology, 55*(5), 331–347.

Kruschke, J. K. and Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. In: J. R. Busemeyer, Z. Wang, J. T. Townsend, and A. Eidels (Eds.), The Oxford Handbook of Computational and Mathematical Psychology, pp. 279-299. Oxford, UK: Oxford University Press.

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2013). *The BUGS book: A practical introduction to Bayesian analysis*. Boca Raton, FL: CRC Press.

Lunn, D., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS — A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing, 10*(4), 325–337.

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods, 16*(4), 406–419.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory and Cognition, 13,* 87–108.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003).* Vienna, Austria.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3,* 382–407.

Riefer, D. M., Knapp, B. R., Batchelder, W. H., Bamber, D., & Manifold, V. (2002). Cognitive psychometrics: Assessing storage and retrieval deficits in special populations with multinomial processing tree models. *Psychological Assessment, 14,* 184–201.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review, 12*(4), 573–604.

Rouder, J. N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review, 12*(2), 195–223.

Serlin, R. C.,& Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist, 40*(1), 73–83.

Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. G. Keren, & C. Lewis (Eds.) A handbook for data analysis in the behavioral sciences: Methodological issues (pp. 199–228). Hillsdale, NJ: Erlbaum.

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E. J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science, 32*(8), 1248–1284.

Smith, M. C., & Thelen, M. H. (1984). Development and validation of a test for bulimia. *Journal of Consulting and Clinical Psychology, 52,* 863–872.

Spiegelhalter, D. J., Freedman, L. S., & Parmar, M. K. B. (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society. Series A, 157,* 357–416.

Stan Development Team. (2012). *Stan: A C++ library for probability and sampling, version 1.1. Retrieved from* http://mc-stan.org/citations.html

Vanpaemel, W. (2009). BayesGCM: Software for Bayesian inference with the generalized context model. *Behavior Research Methods, 41*(4), 1111–1120.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology, 54,* 491–498.

Vanpaemel, W., & Lee, M. D. (2012). Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic Bulletin & Review, 19,* 1047–1056.

Viken, R. J., Treat, T. A., Nosofsky, R. M., McFall, R. M., & Palmeri, T, J. (2002). Modeling individual differences in perceptual and attentional processes related to bulimic symptoms. *Journal of Abnormal Psychology, 111,* 598–609.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14*(5), 779–804.

Wetzels, R., Raaijmakers, J. G. W., Jakab, E., & Wagenmakers, E. J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian *t* test. *Psychonomic Bulletin & Review, 16*(4), 752–760.

Kruschke, J. K. and Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. In: J. R. Busemeyer, Z. Wang, J. T. Townsend, and A. Eidels (Eds.), The Oxford Handbook of Computational and Mathematical Psychology, pp. 279-299. Oxford, UK: Oxford University Press.