This is chapter 4: General frameworks of meta-analysis (pp. 33-54) from

Schulze, R. (2004). *Meta-analysis: A comparison of approaches*. Hogrefe & Huber.

4

General Frameworks of Meta-Analysis

As an example, suppose a group of researchers has succeeded in collecting all empirical studies judged as relevant in a field they are interested in. Recognizing the shortcomings of narratively summing up the collected evidence and confronted with a large amount of empirical evidence, they are interested in statistical methods to quantitatively aggregate the effect sizes extracted from the study reports. Before the researchers turn to specific computational procedures of conducting a meta-analysis, to be described in Chapter 5, they might first consider the following questions:

1. Are there good reasons, theoretically or based on previous evidence, to assume that only *one* universe effect size is underlying all studies? That is, do all studies estimate exactly the same effect?
2. What kind of inference is intended? Should generalization from the results pertain to all potential studies in a field of interest, or should interpretations be restricted to the kind of studies in the collected sample?
3. If studies are not all assumed to estimate the same effect size,
    - are there any theoretically assumed predictors which correspond to observed characteristics of the collected studies to explicitly model potential effect size differences and/or
    - are there potential differences in universe effect sizes that are due to unobserved (latent) variables?

In essence, by answering these questions, the researchers are making a decision between models to be applied to the observed effect size data. Such decisions and arguments to substantiate them have not always been made explicit in published meta-analyses. Often, the choice of a model has been made implicitly by the choice of an approach to meta-analysis. For example, the researchers may turn to one of several textbooks on meta-analysis and apply the

computational procedures outlined in there before considering and answering all the questions outlined above. Unfortunately, not all available textbook resources are explicit with reference to the statistical models implied by the procedures described therein (e.g., Wolf, 1986; but see in contrast, Hedges & Olkin, 1985). Approaches are intimately tied to statistical models, so that the choice of an approach is also the choice of a model. The present section is intended to clarify the basic characteristics of models in meta-analysis. This will provide the framework to classify the specific approaches presented in the subsequent section.

It is important to recognize that *definite* answers to the questions presented above cannot be given on the sole basis of any form of data analysis. The choice of a model has to be made at a conceptual level (Hedges, 1994b; Hedges & Vevea, 1998). This becomes most evident, for example, by considering the second question: What kind of inference is intended? This question can only be answered as a result of careful consideration of the object of inference. On the other hand, there are data-analytical procedures providing some indication of the tenability of a model by way of testing some of its assumptions. In the following sections, such procedures will be presented and their performance under different models will be evaluated on the basis of results of a Monte Carlo study to be presented in Chapter 8.

With respect to the choice of a model, meta-analysis is not at all different from other familiar statistical techniques. Estimation of the parameters of a model is always done by assuming a certain model beforehand, implicitly or explicitly. Structural equation modeling, which has become a very popular statistical technique in practice in recent years, is a prototypical example where one has to choose a model before estimation can be done. However, not all data-analytical techniques force the user to specify or choose between models. Meta-analysis as practised in the field of psychology seems to have become one of these types of data-analytical tools, where decisions of a user are more focused on the choice between sets of computational procedures rather than models.

The question at this point is what kind of models there are available in meta-analysis and which meta-analytical approach corresponds to what kind of model. The following sections are intended to answer these questions. The presentation will thereby be kept more general in comparison to the subsequent presentation of specific approaches (see also Shadish & Haddock, 1994). Although presentation will be focused on the correlation coefficient as an effect size there is no need to restrict the treatment of the subject at this point. Keeping the general perspective in mind, it will be much easier to recognize the similarities and differences of the meta-analytical approaches, and their statistical procedures in particular, to be presented in Chapter 5.

4.1 FIXED EFFECTS MODEL

The fixed-effects model (FE) can still be regarded as the most frequently assumed model in practice. An often stated basic assumption of the FE model is represented in the first question to the researcher in the above list: Are there good reasons to assume a universe effect size that is common to all studies? If the answer to this question is "yes", then the researcher assumes that all observed effect sizes are estimates of a single parameter. The fixed effects model is appropriate for this case.

Let θ denote the universe effect size measure of interest and suppose there are k independent observed effect sizes. This may be the case when an experiment is replicated 10 times ($k = 10$) and each experiment is conducted by a different researcher, in a different place, and so forth, so that all results can be considered as independent. The differences between studies (researcher, place, measurement instruments, etc.) are considered to be minor or negligible in the sense that they do not exert any systematic influence on the research results. The experiments can also be called *strict replications* here. Though such strict replications are only rarely or never conducted in the social sciences, they are assumed for matters of convenience in the presentation at this point.

Furthermore, suppose there is *one* effect size θ giving rise to all effect size estimates. This is a case where effect sizes are often called *homogeneous*, because they all are assumed to represent the same parameter of interest.

However, in general each of the ten replications will report a different observed effect size, so there is a nonzero variance of observed effect sizes. One important question to be answered is how such differences may arise. In the FE model, differences between reported effect sizes are ordinarily conceived as resulting only from sampling error, and sampling error results from different person sampling in the studies. The variance of the observed effect sizes, however, is assumed *not* to be caused by substantive differences between studies, like differences in treatment nuances, validity of measurement instruments, and so forth. This is a very strong assumption for which usage of the FE model has been heavily criticized in recent years (Erez et al., 1996; Hunter & Schmidt, 2000; National Research Council, 1992). As a consequence of the assumptions, one would expect the conduct of 10 more studies of the same type as the first ten studies to result in different estimates of θ only because of varying samples of participants.

The observed effect size measures will be denoted as T_i ($i = 1, \ldots, k$). Usually, different studies also have a different number of participants so that the estimates vary in precision[1] of estimating the parameter θ. The variance of each

[1]In a strict statistical sense estimates do not vary in precision but only estimators do. Hence, one could also conceive each observed estimate as a realization from a different estimator when n is different between studies and the precision of the estimator depends on n. However, in the present context it may be confusing to use the term estimator when all effect size measures are of the same family. As a consequence, the term estimate will be used in what follows.

effect size estimate T_i will be denoted as ν_i and is a measure of this precision. The crucial point is that the estimates might differ in their precision of estimation though they all estimate the *same* constant θ (i.e., $\theta_1 = \theta_2 = \ldots = \theta_k = \theta$). In other words, the universe effect size is *fixed* for all studies.

In order to form a precise pooled estimate based on the observed effect sizes, it seems natural to consider the so-called pooled estimator in the FE model

$$\hat{\theta} = \frac{\sum_{i=1}^{k} w_i T_i}{\sum_{i=1}^{k} w_i}. \tag{4.1}$$

This is also often called the *mean effect size (estimate)*. The connotation implied by this label is that $\hat{\theta}$ is a weighted mean (with weights w_i) of the observed effect size estimates. When all observed effect size estimates are unbiased, then $\hat{\theta}$ is also unbiased. As already shown in the previous chapter, not all measures of effect size of interest are indeed unbiased.

The remaining question is what specific weights are to be inserted in Equation 4.1. From a statistical point of view, the optimal weights are the reciprocals of the variances of the estimates, because they minimize the variance of the pooled estimate $\hat{\theta}$ (for a proof, see Böhning, 2000, pp. 96–97).[2] Therefore, the optimal weights are given by

$$w_i = \frac{1}{\nu_i}. \tag{4.2}$$

Intuitively, these weights also make sense, since they give the largest weight to the most precise estimate (i.e., with smallest ν_i). Since the meta-analytical approaches to be presented in Chapter 5 differ with respect to the choice of weights, a more detailed discussion is postponed to the presentation of the approaches. Note, however, that variances used to compute the weights are usually unknown and have to be estimated. Ordinarily, an estimate for this variance $\hat{\nu}_i$ is available and plugged into Equation 4.2.

No distributional assumptions have been made up to this point. For the next step of inference based on the estimates of θ, it is often assumed that the T_i are normally distributed (e.g., Hedges & Vevea, 1998). However, this is not a necessary assumption to show that

$$\nu_{\hat{\theta}} = \frac{1}{\sum_{i=1}^{k} w_i}$$

[2]It is noteworthy that a justification of the weights can also be given by the maximum likelihood method (see Böhning, 2000, pp. 101-103).

gives the variance of the pooled estimate (see Böhning, 2000), provided every T_i is unbiased. Additionally drawing on the central limit theorem, it is possible to construct confidence intervals via

$$\theta_L = \hat{\theta} - g_\alpha \sqrt{v_{\hat{\theta}}}$$
$$\theta_U = \hat{\theta} + g_\alpha \sqrt{v_{\hat{\theta}}}\,,$$

where g_α denotes the critical value for a prespecified α-level from a standard normal distribution to construct two-sided confidence intervals.[3] The index "L" designates the lower limit and "U" the upper limit of the interval, respectively.

In addition to the construction of a confidence interval, the null hypothesis $\theta = 0$ can also be tested by using $v_{\hat{\theta}}$, so that

$$g = \frac{\hat{\theta}}{\sqrt{v_{\hat{\theta}}}}$$

provides a g-value to be compared with a critical value from the standard normal distribution for a prespecified level of α.

As a last step in the FE model, one can test the basic assumption of equal universe effect sizes underlying all studies by computing the following statistic

$$Q = \sum_{i=1}^{k} \frac{\left(T_i - \hat{\theta}\right)^2}{v_i}.$$

Essentially, this is the sum of squared standard normal values, which follows a χ^2-distribution with $k-1$ degrees of freedom when the null hypothesis of equal universe effect sizes for all k estimates is true. Hence, by comparing the value of Q with the respective critical value from a χ^2_{k-1} distribution, one tests whether the assumption of equal universe effect sizes for all studies holds.[4] This makes the computation of the Q-statistic a very important step in the application of the FE model. When the test result is significant, one is forced to reject the null hypothesis, and this amounts to rejecting the tenability of the FE model. One of the consequences of such a result is that the mean effect size estimate $\hat{\theta}$ has no simple interpretation anymore within the framework of the FE model as presented up to this point.

Of course, it is still the weighted mean of observed effect sizes but the parameter to be estimated is not a single universe effect size constant for all studies. Instead, one is forced to switch to a different model which incorporates differences in universe effect sizes between k studies.

[3]The unusual symbol g_α is used here to avoid confusion with the values of Fisher-z which play a prominent role in the present book.
[4]For convenience, such tests will be labeled as *Q-tests*.

One possibility to deal with the result of a significant Q-statistic is to build subgroups of studies that are assumed to be homogeneous in the sense of the basic assumption of the FE model and compute as many estimates of mean effect sizes as there are groups. Subgrouping in such a procedure can be based on coded characteristics of the studies, for example. These characteristics may be suggested by theoretical reflections or may also be methodological features of the studies (e.g., experimental vs. quasi-experimental studies). In any case, the pursued aim of subgrouping is to find groups that satisfy the assumption of homogeneity in the FE model. A more efficient procedure than subgrouping would be to fit a categorial or continuous (linear) model to the effect size measures as proposed by Hedges (1982a, 1982b, 1994a). In type, these procedures are akin to familiar techniques such as the general linear model (e.g., ANOVA and regression models). A very general framework for this type of analyses is provided by hierarchical linear models, which will be introduced in Section 4.4.

As is well-known, there are also fixed effects models in ANOVA. Indeed, fixed effects models in ANOVA and in meta-analysis are analogous in the sense that the parameter to be estimated is conceived as fixed instead of being a random variable as in the model to be presented next (see, e.g., Scheffé, 1959/1999; for details on the analogy between ANOVA and meta-analysis, see Hedges & Vevea, 1998). It is important to recognize at this point that although the starting assumption of a universe effect size equal for all k studies is rejected, the fixed effects model can still apply.

Nevertheless, the analogy to ANOVA models suggests an interpretational consequence pertaining to the pooled estimate $\hat{\theta}$. Just like in ANOVA, it now has to be interpreted as an estimate of the grand mean of the observed effect sizes. Against this background, the assumption of equal universe effect sizes stated at the outset can be considered as a special case of ANOVA where a factor *study* with k levels has no effect.

Interpretation of results from inferential procedures as outlined above also have to be refined in this model. They now relate to the grand mean built on the basis of a set of k studies which differ in universe effect sizes. The differences between universe effect sizes are now modeled and are considered to be constant (fixed) over replications. If, for example, a meta-analysis is used to aggregate results from ten studies with a certain grand mean, then another set of ten studies must estimate the same grand mean. In this situation one can think of replicating sets of studies with the same grand mean. Hence, inference relates to a *universe* of studies that is characterized by the grand mean to be estimated. The term universe (of studies) is used here again, to underscore the different level of sampling in comparison to primary studies (see also Chapter 2). To reiterate, a first level of sampling can be considered as sampling of persons in the studies, so that there is a population of persons. The second level is considered as sampling of studies form a universe of studies.

Another possibility to deal with the result of a significant test result for the Q-statistic can be to completely give up the fixed effects assumptions and switch to the random effects model to be presented next.

4.2 RANDOM EFFECTS MODEL

The main difference between the FE model and the random effects model (RE) in meta-analysis is the introduction of a random variable Θ instead of an effect in the universe of studies that is conceived as constant (Hedges, 1983b; Raudenbush, 1994). The objects of main focus in RE meta-analyses are the expected value μ_Θ and the variance σ^2_Θ of the random variable Θ. In comparison to the FE model, the expected value of Θ replaces θ as the mean effect size. The variance of Θ is a new object of interest that has no counterpart in FE models. Hence, it is acknowledged in RE models at the outset that universe effect sizes may vary between studies. It is easily seen from this conceptualization that the FE model can also be viewed as a special case of the RE model where the variance of the universe effect sizes is zero and the expected value of Θ and the effect size θ in the FE model coincide.

As a consequence of the model assumption, the variance of observed effect sizes is not only explained by sampling error of persons in studies as was the case in the FE model, but also by *true variability* of studies in meta-analyses. That is, variance of effect size measures is decomposed into two components

$$\sigma^2_{T_i} = \sigma^2_\Theta + \nu_i,$$

where it is assumed that Θ and the error component are independent. The sampling error ν_i of the studies is interpreted as is done in the FE model.

Although there is ordinarily no explicit sampling scheme implied by collecting the studies, it is usually assumed to be a random sampling process. The additional variance component σ^2_Θ — also called heterogeneity variance — introduces an additional source of uncertainty, because apart from sampling n participants at a first level there is also a sampling of k studies with different universe effect sizes at a second level.

The procedures applied in the RE model to estimate a mean effect size first require an estimate of σ^2_Θ. There are different estimators of this variance component that will not be given here, however presentation of specific estimators will be given in the introduction of the refined approaches in Section 5.4 of the following chapter. Assume for the moment that a variance estimate $\hat{\sigma}^2_\Theta$ were available. This estimate is used to compute new weights by

$$w^*_i = \left(\frac{1}{\nu_i} + \hat{\sigma}^2_\Theta\right)^{-1},$$

which are employed in the same way as in the FE model to estimate the mean effect size in the RE model by

$$\hat{\Theta} = \frac{\sum_{i=1}^{k} w^*_i T_i}{\sum_{i=1}^{k} w^*_i}.$$

As can be seen from the weights, results for the mean effect size estimate in the RE model will differ from those in the FE model when the variance estimate $\hat{\sigma}^2_\Theta$ is different from zero, which will generally be the case. Note that $\hat{\sigma}^2_\Theta$ is the same for all k studies so that the effect of the additional component in the weights is to homogenize the weights between studies as compared to the FE model. This also seems plausible since in a situation in which $\hat{\sigma}^2_\Theta$ is much larger in comparison to the v_i this gives a larger impact on the weight to uncertainty due to sampling of studies. In extreme cases where there is practically no estimation error in the individual studies, variability of effect sizes would totally reflect uncertainty due to sampling of studies from the universe. Due to the fact that all studies are ordinarily considered to be equal with respect to sampling from the universe of studies, homogenization is desirable. However, this also makes the estimation of the mean effect size more uncertain and widens the confidence intervals accordingly. This can be seen in the following equations for the construction of confidence intervals

$$\Theta_L = \hat{\Theta} - g_\alpha \sqrt{v_{\hat{\Theta}}}$$
$$\Theta_U = \hat{\Theta} + g_\alpha \sqrt{v_{\hat{\Theta}}}$$

where, again, g_α is the critical value from a standard normal distribution. The estimate of the variance of $\hat{\Theta}$ is denoted by $v_{\hat{\Theta}}$ and given by the reciprocal of the sum of weights

$$v_{\hat{\Theta}} = \frac{1}{\sum_{i=1}^{k} w_j^*}.$$

In the same fashion as in the FE model but with the new weights, a significance test for the hypothesis $\Theta = 0$ can also be performed by

$$g = \frac{\hat{\Theta}}{\sqrt{v_{\hat{\Theta}}}}.$$

As can easily be seen by considering computation of the weights in the RE model, the tests are — ceteris paribus — generally less powerful than those of the FE model. This is due to the additional component $\hat{\sigma}^2_\Theta$ which makes the weights larger, and as a consequence, standard errors $v_{\hat{\Theta}}$ also become larger.

The estimates $\hat{\Theta}$ are always clear to interpret in the RE model. They represent estimates of the expected value of the distribution of universe effect sizes. This is an important point to note since the distribution of effect sizes in the universe of studies represents the distribution of all possible studies. The universe comprises the k studies in a meta-analysis as a sample but also all other studies that could not be retrieved (see Hedges & Vevea, 1998). This suggests a very attractive interpretation of the mean effect size estimate in RE models, namely that the effect size estimate may be generalized to an entire research domain. This is one of the reasons why some authors have argued strongly in favor of the application of the RE instead of FE models (e.g., Hunter & Schmidt,

2000). In other methodological areas in psychology, like generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972), there is also an analogous transition in models where the RE model is strongly favored.

However, there is always some ambiguity left in interpretation when the sampling process is somewhat obscure as will mostly be the case in applications of meta-analysis. A random sampling process would require the specification of the whole universe of studies and a procedure that guarantees a random sample of k studies from this universe. This is not feasible in practice and may represent a critical point for the application of RE models. In a similar vein, some authors have noted that the assessment or decision as to whether study samples are indeed representative for an entire research domain is not an easy task, if possible at all (Kavale, 1995). Yet this is not a problem specifically pertaining to meta-analysis but also arises in ordinary research practice in psychology or other fields where random samples are scarcely available. On the other hand, random sampling is considered not to be a necessary prerequisite in general for valid interpretations by some authors (e.g., Frick, 1998). Furthermore, a Bayesian perspective on the research problem in meta-analysis also does not necessitate a formal random sampling procedure for the justification of random effects (Raudenbush, 1994).

In addition, when not many studies are available in a field of interest generalization to a whole domain of research may be unfounded or at least risky because few studies are scarcely representative for a universe of studies. Furthermore, problems arise also in the application of RE models to a set of only few studies with respect to estimation of the heterogeneity variance σ^2_Θ (Raudenbush, 1994; see also Hunter & Schmidt, 1990, who discuss such issues under the heading of *second-order sampling error*). It may be more sensible in such cases to restrict interpretation only to studies like those in the sample as is done with the FE model. Therefore, it is of great interest how applications of the RE model perform in situations with very few studies which is one of the aims that will be pursued in the empirical part of this book.

Unfortunately, there seems to be considerable confusion as how to conceptualize and interpret the random effects model of meta-analysis. For example, Erez et al. (1996) draw a distinction between the fixed and random effects model in a way that the fixed effects model is interpreted as an intercept-only regression model, whereas the random effects model is regarded as a regression where the heterogeneity of observed effect sizes is additionally accounted for by covariates (Erez et al., 1996, p. 278). The difference between the FE and RE model, however, is not one of differences in predictors in regression models but whether universe effect sizes are conceived as random variables or not. In both models it is possible to apply linear models for the explanation of variation in study findings with any desired set of predictors as long as the basic assumptions of the models are met.

As already noted, the homogeneity test based on the Q-statistic in the FE model is often used in practice to make a decision between the random and the fixed effects model. Although this decision does not require such a test, the decision is often made conditionally on the result of the test. Such a proce-

dure is also called the *conditionally random effects procedure*. This hybrid procedure has been reported to have properties in between FE and RE procedures with respect to test results (Hedges & Vevea, 1998). As has been outlined in this section, there are important differences in interpretation associated with the choice of a model. Therefore, it seems reasonable to require the Q-test to perform quite well as one of the most important decisions in meta-analysis hinges on its results. The present study will also present an empirical evaluation of the Q-test as used in various approaches to assess its quality (see also Alexander, Scozarro, & Borodkin, 1989; Cornwell, 1993; Field, 2001; Hardy & Thompson, 1998; Harwell, 1997; Hartung, Argaç, & Makambi, 2003; Sánchez-Meca & Marín-Martínez, 1997).

4.3 MIXTURE MODELS

Mixture models provide a very general framework for the meta-analytic situation that embrace and extend the fixed and random effects models presented in the previous two sections. Since mixture analyses are not part of the Monte Carlo procedures to be presented in later chapters, only a brief sketch of the main characteristics is given here. The concepts introduced in this section will nevertheless be taken up in later sections because they provide a very concise way to describe the meta-analytical situation in a well-founded statistical theory. For an in-depth treatment of the subject with application to meta-analysis the reader is referred to the work of Böhning (2000) and also to one of the first applications of these methods to meta-analysis in psychology by Thomas (1989a, 1989b, 1990b). Because the present study is mainly occupied with the application of meta-analysis to correlational data, the following presentation will be given with the correlation coefficient as effect size data.

Suppose again, there are $k = 10$ studies given and each of the ten studies reports a correlation coefficient r_i for two variables that are bivariate normal in distribution. Now the following concepts and notation are introduced. Observed correlations are regarded as realizations of random variables denoted by R_i with a certain n_i per study and universe correlation ρ_j. For matters of convenience, it is assumed that all n_i are equal[5] and can be denoted by n. The index j is used to indicate potentially different universe correlations for a set of i correlations. That is, there may be subsets of the i studies with different universe correlations, for example. In mixture models, such j universe parameters — universe correlations in the present case — are also called *components*. The total number of components is denoted by c, so that $j = 1, \ldots, c$.

What the meta-analyst wants to understand, explain, and model, is how the distribution of observed correlation coefficients arises. If there is only one $\rho_j = \rho$ common to all studies, a homogeneous case is given. In mixture models

[5]It would not be difficult to conceive the number of participants also as a random variable. However, this would not add much to understanding the concepts here and there is no loss in generality by assuming equal n.

this is also called a case with one component. The distributions of the R_i only differ when the sample sizes n_i of the studies are different. Otherwise, all variables have the same distribution, characterized by a probability density function $f(r;\rho,n)$. In this case, knowledge of ρ and n suffices to characterize the sampling distribution of the observed correlation coefficients.

Now suppose, two components with $\rho_1 \neq \rho_2$ and therefore a heterogeneous case is given. As mentioned in the previous sections, such a situation could be modeled by procedures of the general linear model when it is known for each of the k studies which of the two ρ_j is underlying each study. Assume such knowledge is not available to the meta-analyst and membership of the observed correlations to the different components can be said to be unobserved or latent. In this situation, the distributions of the R_i differ only because of the different ρ_j. The ρ_j can now themselves be considered as realizations of a random variable P (large Greek Rho)[6]. The distribution of P is called the *mixing distribution* in the present context and is not yet specified. For the present case of only two different components ρ_1 and ρ_2, the distribution of P is characterized by the two components and the according weights λ_j. The weights give the probability of belonging to the jth component and therefore conform to the usual constraints $\lambda_j \geq 0$ and $\sum_{j=1}^{c} \lambda_j = 1$ when there are c components (in the present example, there are only two).

Under these conditions, the correlation coefficient R_i of interest in study i can be said to have a *conditional density* denoted by $f\left(r_i|\mathrm{P}=\rho_j,n\right)$. That is, given the universe effect size parameter ρ_j and the number of participants per study n, the correlation coefficient has a density as given in Equation 3.1 (see page 21). For purposes of illustration, assume that for the two components in the example all studies have equal probability of belonging to one of the components. That is, $\lambda_1 = \lambda_2 = .50$. For this example, the unconditional density of R, the variable representing all observed correlation coefficients r, is given by

$$f(r|n) = .50 \times f(r;\mathrm{P}=\rho_1,n) + .50 \times f(r;\mathrm{P}=\rho_2,n),$$

and for the more general case of c components the unconditional density is

$$f(r|n) = \sum_{j=1}^{c} \lambda_j \times f(r;\mathrm{P}=\rho_j,n).$$

This is also the density of the so-called *mixture distribution* with *kernel* $f(r;\mathrm{P}=\rho_j,n)$ for the present case. Of course, the kernel of the mixture distribution depends on the effect size under investigation when mixture models are applied in the context of meta-analysis.

[6]Not to be confused with the symbol for probability $\mathcal{P}$ used in the following. In any case, it will also always be clear from the context which symbol is used.

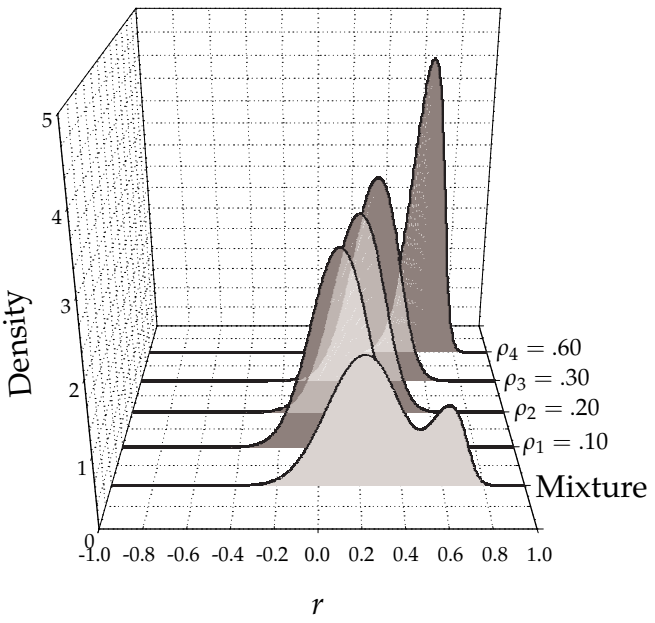

Figure 4.1 Example of a mixture distribution with four components, $n = 50$.

As a further example, consider a situation with $n = 50$ where the distribution of P is uniform on the four points $\rho_1 = .10$, $\rho_2 = .20$, $\rho_3 = .30$, and $\rho_4 = .60$. Then a situation like the one depicted in Figure 4.1 is given.

The four darkly shaded densities are the conditional densities for each of the components, with a fixed n of 50. The resulting mixture distribution filled in light grey is depicted in the front and illustrates the density of R in this situation. In meta-analysis, a number of k correlation coefficients are given which are considered as arising from the mixture density given in Figure 4.1. That is, the mixture distribution is similar in shape to what one would expect as a frequency distribution of k observed correlations in a meta-analysis (given the four components and a fixed n of 50).

As already mentioned, the number of components is usually unknown so that it has to be estimated along with the component weights. Conventionally, this is done by maximum likelihood estimation but details on estimation and algorithms will not be presented here (see Böhning, 2000).

The attractive options offered to the meta-analyst by an application of mixture models are manifold. Mixture models provide a general and flexible framework of conceptualizing as well as statistically modeling the object of interest in meta-analysis, namely the distribution of observed effect size measures. Furthermore, procedures to estimate the number of components as well as their weights are offered. This makes it possible to address the problem of heterogeneity of effect sizes even after attempts to apply linear models with observed variables have been undertaken. When the number of components and their weights are estimated, it is also possible to classify the k studies under investigation by posterior Bayes classification (see Böhning, 2000). The fit

of the model applied to effect size data can also be assessed to give an impression of how well the estimated parameters serve to explain heterogeneity.

Of course, the application of mixture models to effect size data does not guarantee that the user can easily interpret the composition of the components suggested. Interpretation of results requires theorizing as well as speculation about the nature of the latent variable. Replications and further research may well be indicated to support or to question interpretation of results from mixture analysis.

Evaluations of an early mixture approach to meta-analytic databases by Thomas (1989b, 1990b) on the basis of Monte Carlo study results were quite encouraging (see Law, 1992). In several situations of Law's study, the procedures proved to be quite accurate with respect to estimation of the weights and the actual values of ρ_j. However, in identifying the proper number of components there seemed to be room for enhancement of the procedures. Given that improved algorithms and procedures have become available in recent years, updated and more in-depth evaluations of the procedures seem to be desirable.

To conclude, as with many statistical techniques newly introduced to a field of application, mixture models involve relatively complicated procedures and estimation is by far not as easily done as with the procedures outlined for the FE and RE models. However, in the case of mixture models there are easy-to-use programs available so that estimation is feasible in practice and therefore not really much more complicated than with all other models (Böhning, Schlattmann, & Lindzey, 1992; Schlattmann, Malzahn, & Böhning, 2003).

4.4 HIERARCHICAL LINEAR MODELS

In addition to the more standard FE and RE models and the more advanced mixture distribution analysis presented in this chapter, there are other models available as well. These will not be treated in detail, but at least a rough idea of their basics of conceptualization will certainly help in gaining a deeper understanding of meta-analysis and potential modeling approaches. This section describes the HLM approach. For a comprehensive overview of models, (estimation) methods, and issues in HLM that also includes meta-analysis as a special case, the reader may consult the book by Raudenbush and Bryk (2002). A more focused and succinct presentation on multilevel models for meta-analysis is given by Hox and de Leeuw (2003), for example.

Hierarchical linear models (HLM) are a very general class of models that may be applied not only in meta-analysis, but in a very large number of situations, all of which are characterized by different levels of data. The lowest level of data in HLM is ordinarily the individual units level, that is, persons in an observational or experimental study, for example (Level 0; see Figure 2.1). As a result of primary analyses, some estimates for parameters of interest are obtained and these are considered to constitute another level of data (Level 1). As conceptually outlined in the previous sections, such estimates are the data for

meta-analysis. If modeling of the study parameters (θ_i) is of interest, then we have yet another level of data. This is the case, for example, in mixture modeling as presented in the previous section, where study parameters are thought to arise from a mixing distribution. In sum, it is important to recognize and differentiate levels of data as conceptualized in HLM.

Yet, the situation of meta-analysis is somewhat special from the perspective of HLM. The data of individual units are not available and if they were, one would most probably try to conduct secondary analyses or more specifically a three-level analysis in HLM. The first level of interest in meta-analysis is therefore at the study level and modeling takes place at a second level in order to explain potential heterogeneity of effects (i.e., $\sigma^2_\Theta \neq 0$), for example.

To explain how the meta-analytical situation is modeled with HLM, consider once more the situation a meta-analyst is confronted with. There are a number of k study results, extracted from the literature on a certain research question, and the task is to summarize them in a theoretically sound and — for the research question — appropriate way. The following equation specifies a model for the individual effect size data of the ith study, that is, a so-called *within-studies* or Level 1 model by

$$T_i = \theta_i + e_i. \tag{4.3}$$

The observed effect size is a realization of T_i for the ith study. It is conceived as the sum of the corresponding universe parameter θ_i and an error component denoted by e_i. The error component represents random fluctuations, whereas the universe parameter θ_i is a constant per study and hence specific for every study i. As an alternative, one might as well assume $\theta_1 = \cdots = \theta_k = \theta$ as is done in FE models. This additional assumption makes the FE model a special case of HLM. The error component is ordinarily assumed to be normally distributed with expected value of zero, that is, $e_i \sim \mathcal{N}(0, \nu_i)$. The variance of the error component ν_i can therefore be identified as error variance of the estimator T and is assumed to be *known* in HLM of meta-analysis. This latter assumption results from the situation given in meta-analysis, where the available data have to be gained from research reports and original data at the individual level are not available.

For the case of correlational data, Equation 4.3 may be stated as $r_i = \rho_i + e_i$. What is already known from the previous chapters is that for correlations as effect sizes, the assumption of a normal distribution for the error component is not tenable for sample sizes less than approximately 500. For this reason, the correlation coefficient is ordinarily transformed into z-space by the Fisher-z transformation for which the assumption of normally distributed errors may be reasonable even for modest sample sizes. In addition, assuming the error variance to be known is also well-founded in z-space since it only depends on n_i and no estimation is needed. Note, however, that ν_i would have to be estimated for correlations and the error component variance involves the universe parameter (see Equation 3.6 on page 26). Furthermore, correlations are biased and the assumption $E(e_i) = 0$ is not correct in a strict sense, though the bias

may be negligible, especially for ρ_i close to zero (see Chapter 3). The same is true for values resulting from the Fisher-z transformation. As will be shown in later chapters, however, this transformation has some undesirable properties making its use a problematic feature in HLM and meta-analysis in general.

In addition to the Level 1 model, the following Equation is of importance in HLM. It specifies the Level 2 — or *between studies* — model as

$$\theta_i = \gamma_0 + \gamma_1 X_{1i} + \cdots + \gamma_L X_{Li} + u_i. \tag{4.4}$$

The linear model stated in this equation includes a set of L regressors X all of which are considered to be *observed* study characteristics in meta-analysis. Examples for such variables include methodological quality scores and other attributes coded in step 3 of a meta-analysis (see Chapter 2), like intensity or duration of an experimental treatment, type of measurement instruments used, and so forth, which are more of substantive interest. The parameters in the equation are the intercept γ_0 and the weights for the regressors $\gamma_1, \ldots, \gamma_L$. These components of Equation 4.4 represent the explanatory part for the variability in θ_i. Additionally, there is a random effect component for each study denoted by u_i. This random effect represents each study's universe parameter θ_i deviation from the value predicted by the explanatory part of the model. The random effect component is ordinarily assumed to be normally distributed as $U \sim \mathcal{N}(0, \sigma_U^2)$ in HLM. This makes clear that the study parameters θ_i are conceived as realizations of a random variable θ. Due to the fact that the model includes fixed effects (the regressors) and a random effect (u_i) the model is referred to as a mixed model.

Substituting Equation 4.4 in Equation 4.3 results in

$$T_i = \gamma_0 + \sum_l \gamma_l X_{li} + u_i + e_i. \tag{4.5}$$

In this equation it becomes clear how variability of the observed effect size measures T_i is decomposed in HLM. There is variance explained by the study characteristics, there is residual variability due to a random effect, and also variability due to sampling error. HLM is quite an attractive model for meta-analysis that goes beyond the more standard models of fixed and random effects as outlined in Sections 4.1 and 4.2 by incorporating explanatory variables. It includes, however, these more popular models as special cases. The generality of HLM is thus recognized by considering some special cases of Equation 4.5.

First, the FE model without explanatory variables was already shown to be a special case. Second, consider Equation 4.5 without a random effect u_i. This basically is the fixed effects model in meta-analysis with regressors as described by Hedges and Olkin (1985). Note that in such models there is, of course, the supposition of variability in the T_i, but it is assumed to be explained by the regressors so that only variability due to the error component remains. Hence, σ_U^2 is assumed to be zero. One of the important features of HLM is that such assumptions are testable. HLM therefore offers statistical tests in meta-

analysis — in this specific case akin to the Q-test — to test critical assumptions. Third, imagine there were no explanatory variables in Equation 4.5, so that

$$T_i = \gamma_0 + u_i + e_i.$$

In this case, the model is equivalent to the RE model as presented in Section 4.2. The intercept γ_0 represents the mean effect size across all studies, the variability of σ_U^2 would correspond to σ_Θ^2, and the variability of e_i is v_i. As alluded to before, tests and the construction of confidence intervals are possible by using HLM to analyze a meta-analytic database.

However, HLM does not include all models presented in this chapter as special cases. An important exception are mixture models. Although both models aim at explaining potential heterogeneity of effect sizes, HLM incorporates observed explanatory variables, whereas in mixture models such variables are considered as latent. Hence, both models should be considered as complementary rather than competing.

Apart from their theoretical attractiveness, how well do HLM perform in comparison to the more simple and much more popular standard FE and RE models? Since different — likelihood-based — estimation algorithms are used in HLM, it can not be taken for granted that they lead to the same or better results as compared to standard models. Available Monte Carlo studies focusing on the standardized mean difference as an effect size show that HLM methods compare quite favorably under some simulated conditions. Van den Noortgate and Onghena (2003) have made such a comparison and showed that HLM lead to very similar results vis-à-vis RE models for parameter estimates, for example. Interestingly, they also pointed to some deficiencies in the testing procedures, for instance, and concluded that HLM procedures do not unequivocally lead to better results in comparison to standard models. Nevertheless, this does not belittle the virtue of model generality of HLM.

Finally, as important extensions of the basic HLM for meta-analysis, there are multivariate models available which enable the meta-analyst to deal with the otherwise difficult situation of multiple effect sizes per study, a case quite often encountered in practice. Another important problem in meta-analysis, namely missing data for regressor variables, can also be handled in a statistically sound way with HLM. All of these extensions are well beyond the scope of interest in the present context. In addition to the book by Raudenbush and Bryk (2002), the interested reader is referred to Kalaian and Raudenbush (1996) for multivariate extensions.

4.5 CLASSES OF SITUATIONS FOR THE APPLICATION OF META-ANALYSIS

The following presentation serves several purposes. It provides a taxonomy of classes of situations that will more clearly elucidate potential forms of distributions in the universe of studies when correlation coefficients are used as effect sizes. Furthermore, the conceptual distinctions to be introduced will also

serve to concentrate the subsequent presentation on some of the members of the classes of situations. Finally, the presentation specifies the distributions in the universe of studies of concern in the third part of the book.

Against the background of the models introduced in the previous sections, several distinct situations, henceforth denoted by $\mathfrak{S}$, can be identified. The term *situation* is used throughout the present and the following chapters in a generic sense to indicate distinct classes of universe effect size distributions. In analogy to the presentation of mixture models in Section 4.3, the universe effect sizes of the studies to be aggregated can be regarded as realizations ρ of a random variable P. The expected value of this variable will be denoted by $E(\mathrm{P}) = \mu_\rho$ and its variance by σ_ρ^2. Suppose there is a total number of k studies, so that we have $\rho_1, \ldots, \rho_k$. Then the situations to be described in the following two paragraphs will be distinguished by the form and parameters of the distribution of P, that is, the parent or mixing distribution. Two broad types of classes can be differentiated here: discrete and continuous mixing distributions.

Discrete Distributions. There is one important special case among the discrete distributions that defines the first situation $\mathfrak{S}_1$, namely a one-point distribution with probability mass 1 at the point of a single ρ_0 in $[-1, 1]$. That is,

$$\mathcal{P}(\rho) = \begin{cases} 1 & \text{if } \rho = \rho_0, \\ 0 & \text{otherwise} \end{cases}$$

This is the most simple distribution, where the universe of studies is characterized by a *single* constant effect size $\rho = \rho_0$ that gives rise to *all* observed effect sizes. As a consequence, no variation of universe effect sizes is present here (i.e., $\sigma_\rho^2 = 0$), a situation for which the FE model is appropriate. Since all studies are identical with respect to ρ, a *homogeneous* situation is given. To illustrate one instance of $\mathfrak{S}_1$, assume $\rho_0 = .40$. In this situation, the universe parameter for all studies is .40 with probability 1 and the sampling distribution of the observed correlation coefficients r_i is exactly the same for all studies if all studies have the same number of persons n_i, that is, $n_1 = n_2 = \ldots = n_k = n$ (see Figure 4.2).

In the upper panel of Figure 4.2, a graph of the discrete density of the mixing distribution is depicted. The probability mass is concentrated at the point $\rho_0 = .40$ and all other values of the interval from -1 to 1 have zero probability. This universe parameter is underlying all studies so that the density of the observed coefficients is the same for all studies. This density is depicted in the lower panel of Figure 4.2. Here it is assumed that in all studies an n of 50 is given. The conditional density $f(r|\mathrm{P} = \rho_0)$ depicted in the lower panel is — although it looks like a normal distribution at first glance — the exact density given by Hotelling (1953) (see page 21). Accordingly, the distribution of the random variable R_i for the observed effect sizes is fully determined by ρ_0 and the number of participants n in the k studies to be aggregated. In effect, one can argue that there is nothing to differentiate on the level of the universe in

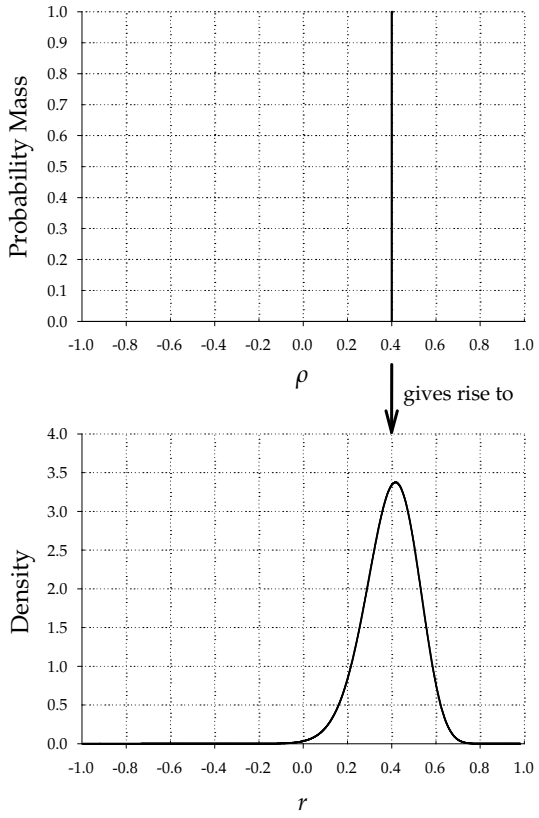

Figure 4.2 Example of $\mathfrak{S}_1$, $\rho_0 = .40$ and $n = 50$.

the upper panel and there is also no need to set sampling distributions apart as long as all the studies have the same n. This is entirely true for a fixed effects model and it will become evident that the differentiations were made for conceptual reasons.

The situation depicted in Figure 4.2 is highly restricted with respect to the distribution of the values in the universe and one might wonder whether $\mathfrak{S}_1$ is relevant at all for the present study. However, as already noted it is actually the most often assumed model in published meta-analyses. Although this assumption is rarely explicitly stated, it is implied by the application of FE methods in meta-analysis as described in Chapter 5. Furthermore, albeit not plausible as a model for most research situations in psychology, there may be cases for which the FE model seems appropriate for theoretical reasons or based on research experience. Even though strict replications — for which the FE model would be a perfectly reasonable model — regrettably are exceptions in the social sciences, there are at least some fields like personnel selection for which a homogeneity "at the level of substantive population parameters" can be assumed on the basis of research experience (Hunter & Schmidt, 2000, p. 276; see also Schmidt et al., 1993).

In the second class of situations with a discrete distribution $\mathfrak{S}_2$, *two* subpopulations are present at the the universe level. They are characterized by two

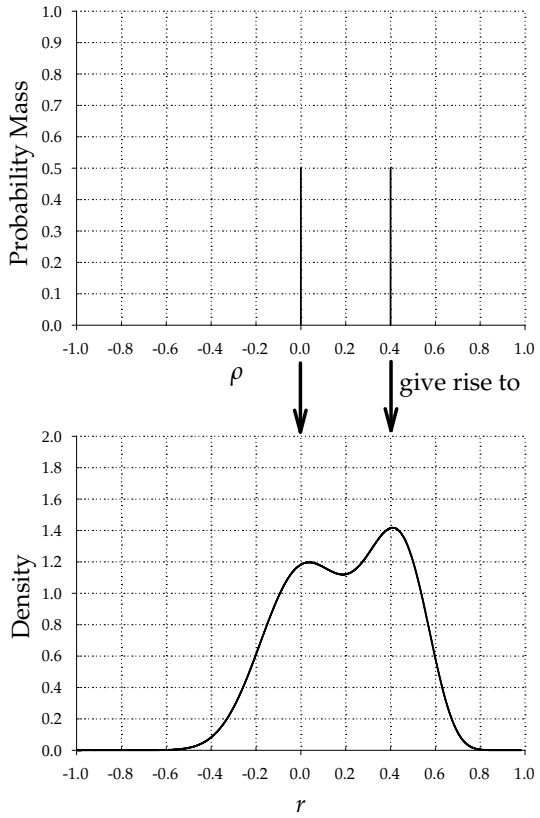

Figure 4.3 Example of $\mathfrak{S}_2$, $\rho_1 = .00$, $\rho_2 = .40$, and $n = 32$.

discrete and distinct values ρ_1 and ρ_2 in $[-1, 1]$. Specifically,

$$\mathcal{P}(\rho) = \begin{cases} .50 & \text{if } \rho = \rho_1, \\ .50 & \text{if } \rho = \rho_2, \\ 0 & \text{otherwise} \end{cases} \tag{4.6}$$

The variance of P is different from zero in this situation but P can only take on two values. This is clearly a heterogeneous case. The following presentation will exclusively be restricted to instances in which $\mathcal{P}(\mathrm{P} = \rho_1) = \mathcal{P}(\mathrm{P} = \rho_2) = .50$, so that $\mathcal{P}(\mathrm{P} = \rho_1) + \mathcal{P}(\mathrm{P} = \rho_2) = 1$, of course. Both values ρ_1 and ρ_2 are therefore equally likely to occur. Hence, it is assumed that the k studies to be aggregated are sampled with equal proportions from one of the two classes, respectively. Of course, different cases of discrete two-point distributions with unequal probability masses can easily be imagined but for convenience the presentation will be restricted to the special case indicated. An example for $\mathfrak{S}_2$ with values $\rho_1 = .00$, $\rho_2 = .40$ and $n = 32$ for the studies in both groups is depicted in Figure 4.3.

The upper panel of Figure 4.3 again shows the distribution of the effect sizes in the universe of studies, now with equal probabilities of .50 for both components. The mixture distribution arising from this mixing distribution is de-

picted in the lower panel of Figure 4.3. Here, it becomes evident how multiple mode or extremely skewed empirical distributions of correlation coefficients may arise in practice. Again, the mixture distribution is derived from the exact density of the correlation coefficient given by Hotelling (1953). Drawing on the notation introduced in Section 4.3, a variable P is given taking on two possible values ρ_1 and ρ_2. It describes the membership of the subgroups in the universe of studies as in Equation 4.6. As usual in mixture distribution analysis, the unconditional density is given by

$$f(r|n) = \mathcal{P}(\rho_1) \times f(r|\rho_1, n) + \mathcal{P}(\rho_2) \times f(r|\rho_2, n)$$

This density is depicted in the lower panel of Figure 4.3 for $n = 32$ in the studies of both classes.

As an interpretation from a substantive viewpoint, the heterogeneous case of $\mathfrak{S}_2$ can be interpreted as corresponding to research situations in which there is an unobserved discrete variable P that moderates the research results. Of course, if one knew about this variable — especially if it could be represented or approximated by observed characteristics of the studies under investigation — efforts to model effect size differences within the framework of HLM or to identify the subgroups by mixture analyses would certainly be indicated. However, it is *not* the aim of the present investigation to evaluate explanatory models[7] in meta-analysis (for a Monte Carlo study on this topic, see Overton, 1998). Instead, it will be assessed how the most often applied methods of meta-analysis perform when data is collected in the heterogeneous situation $\mathfrak{S}_2$. It is argued that it is far from an uncommon situation that a moderator goes unrecognized in a meta-analysis or that estimates of mean effect sizes using the FE model are presented in heterogeneous situations (for a series of examples, see Hunter & Schmidt, 2000).

The question arises in such cases whether an estimate of a mean effect size is sensible at all and if so, how such reported mean effect sizes are to be interpreted. This is not an easy question to answer because it depends on the parameter one intends to estimate and the kind of inference to be made. The presence of heterogeneity per se as in the given situation does not necessarily preclude the reasonable application of fixed effects analysis (Hedges & Vevea, 1998) and the computation of a mean effect size. If one wishes to characterize the study sample with the given characteristics and no further inference is intended, then it is perfectly reasonable to apply fixed effects methods, but the interpretation of test results has to be restricted to studies like those in the sample (see also Section 4.2). The mean effect size that results from applying these procedures is intended to estimate the expected value of the effect size distribution in the universe of studies, just like the grand mean in ANOVA analyses, and has to be interpreted in a similar way in heterogeneous situations. Thus, it

[7]Such regression-type models are also known as "moderator analysis" in the social sciences literature. Most often, such regression-type models do not include a random component (u_i) and can therefore be considered to be a special case of the more general HLM for meta-analysis. These special cases are known as "meta-regression" in the medical literature.

has to be conceived as a mean of potentially very different values of universe effect sizes. Any of such mean effect sizes is therefore ambiguous in the sense that vastly different ρ_1 and ρ_2 might yield the same mean effect size. Nevertheless, though ambiguous, a value of .45 for the relationship between attitudes and behavior, for example, can be considered as informative when the additional assumption that the values in the universe of studies are not very different is tenable. Hence, the question whether such values make sense is not a statistical one but has to be answered by the researcher who applies such procedures. Consider yet another example. If interest lies in the predictive validity of a personnel selection procedure in country A in comparison to country B, then one would synthesize all results from applications in these countries separately and make a comparison of estimated mean validities at this level of aggregation. Of course, there may be differences in validities *within* countries, but these are not of interest for the comparison as long as differences within countries occur equally in both groups.

These remarks are definitely not intended to argue in favor of fixed effects analyses or in any way against the application of explanatory models within HLM, for example. Instead, they only illustrate that an estimate of a mean effect size can indeed make sense in heterogeneous situations like $\mathfrak{S}_2$ in the way just described.

Continuous Distributions. The third class of situations $\mathfrak{S}_3$ is characterized by a continuous distribution of the correlation coefficients in the universe of studies. The realizations of P do not take on any restricted or discrete set of values in the universe but are spread over the entire interval from -1 to 1. The kind of spread is described by a continuous density f. The form of this distribution is ordinarily unknown but it is often assumed to be a normal distribution (Lau et al., 1998). There may be several reasons why the normal distribution is chosen. First, lack of prior knowledge about the exact composition of a presumed myriad of influences that determine the effect sizes in a class of research situations, and arguments in analogy to the central limit theorem let the normal distribution appear as a good guess for the distribution at least. Second, especially in situations where effect sizes that can be shown to have a normal sampling distribution are of concern it seems reasonable, again by way of analogy, to assume the same distribution for the universe effect sizes as for its sampling counterpart. Finally, familiarity with and ease of statistical tractability of the normal distribution also contribute to the fact that it is chosen quite often as the distribution of universe effects sizes. Although none of these reasons is essentially compelling there are no cogent alternatives available in such a state of lack of knowledge.

However, for the present case of correlations as effect size data it would be implausible to assume a normal distribution for the universe correlations, due to the fact that the range of these coefficients is bounded by the values -1 and 1. Especially when high absolute values are of particular interest, the normal distribution would provide invalid values larger than 1 or less than -1. The normal distribution has nevertheless been used in simulation studies

of meta-analyses to generate values of the universe correlation coefficient (e.g., Overton, 1998).

The question arises which continuous distribution might be considered instead of the normal distribution. Several such distributions were considered as candidates which had to — as was the case with the normal distribution — appear as reasonable for the distribution of the effect sizes in the universe. They also had to conform to the requirement of being supported by the interval $[-1, 1]$. The family of beta distributions was finally considered to be the most sensible choice. It was chosen because its parameters can be adjusted to yield a series of very different distributions on the desired interval. The great flexibility of the beta distribution and the ease of its tractability also made it particularly useful for the present purpose (see also Hedges, 1989). Moreover, the parameters of the beta distribution can be chosen so that the distribution is symmetrical at $\rho = 0$ with an increasing skew for larger values of ρ (in absolute terms). Such distributions resemble the sampling distribution of the correlation coefficients r_i (see Section 7.3). To illustrate, Figure 4.4 depicts a series of beta distributions which show the properties just mentioned.

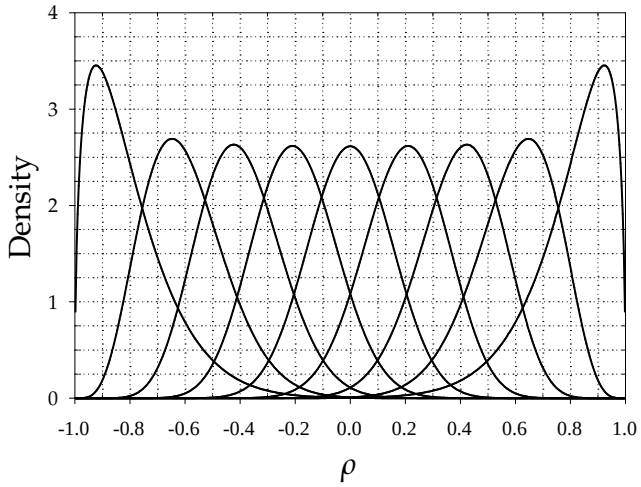

Figure 4.4 Beta-Distributions in $\mathfrak{S}_3$ with varying μ_ρ from $\mu_\rho = -.80$ to $\mu_\rho = .80$ in increments of .20, $\sigma_\rho = .15$ for all distributions.

The parameters of the beta distributions shown in Figure 4.4 were chosen to have different expected values μ_ρ from $-.80$ to .80 in increments of .20 but with a constant standard deviation of $\sigma_\rho = .15$. As is evident from the distributions given, their forms do at least seem plausible for the given range of ρs.

In the Monte Carlo study to be presented, the family of beta distributions will be considered as the distribution of effect sizes in the universe of studies in $\mathfrak{S}_3$. The following theoretical examinations will often abstract from the specific distributional form but sometimes the beta distribution will be used for illustration purposes.