the standard error increases as the number of clusters decreases with fixed sample size. For the higher values of intraclass correlation shown in the graphs, it seems that it would be best to choose enough clusters so that no more than 20 persons are selected within each cluster.

But why, in Figure 20.4, do we think that interclass correlations between 0 and 15% are plausible? To start with, for binary data, the denominator of (20.3) can be reasonably approximated by 0.25 (since $p(1-p) \approx 0.25$ if p is not too close to 0 or 1). Now suppose that the clusters themselves differ in some particular average outcome with a standard error of 0.2—this is a large value of σ_α, with, for example, the percentages of Yes responses in some clusters as low as 0.3 and in others as high as 0.7. The resulting intraclass correlation is $0.2^2/0.25 = 0.16$. If, instead, $\sigma_\alpha = 0.1$ (so that, for example, the average percentage of Yes in clusters varies from approximately 0.4 to 0.6), the intraclass correlation is 0.04. Thus, it seems reasonable to consider correlations ranging from 0 to 5% to 15% as in Figure 20.4.

20.5 Multilevel power calculation using fake-data simulation

Figure 20.5a shows measurements of the immune system (CD4 percentage, transformed to the square root scale to better fit an additive model) taken over a two-year period on a set of HIV-positive children who were not given zinc. The observed noisy time series can be fitted reasonably well by a varying-intercept, varying-slope model of the form, $y_{jt} \sim \mathrm{N}(\alpha_j + \beta_j t, \sigma_y^2)$, where j indexes children, t indexes time, and the data variance represents a combination of measurement errors, short-term variation in CD4 levels, and departures from a linear trend within each child. This model can also be written more generally as $y_i \sim \mathrm{N}(\alpha_{j[i]} + \beta_{j[i]} t_i, \sigma_y^2)$, where i indexes measurements taken at time t_i on person $j[i]$. Here is the result of the quick model fit:

R output

```
lmer(formula = y ~ time + (1 + time | person))
            coef.est coef.se
(Intercept)  4.8      0.2
time        -0.5      0.1
Error terms:
 Groups   Name        Std.Dev. Corr
 person   (Intercept) 1.3
          time        0.7      0.1
 Residual             0.7
# of obs: 369, groups: person, 83
```

Of most interest are the time trends β_j, whose average is estimated at -0.5 with a standard deviation of 0.7 (we thus estimate that most, but not all, of the children have declining CD4 levels during this period). The above display also gives us estimates for the intercepts and the residual standard deviation.

We then fit the model in Bugs to get random simulations of all the parameters. The last three panels of Figure 20.5 show the results: the estimated trend line for each child, a random draw of the set of 83 trend lines, and a random replicated dataset (following the principles of Section 8.3) with measurements at the time points observed for the actual data. The replicated dataset looks generally like the actual data, suggesting that the linear-trend-plus-error model is a reasonable fit.

Modeling a hypothetical treatment effect

We shall use these results to perform a power calculation for a proposed new study of dietary zinc. We would like the study to be large enough that the probability is

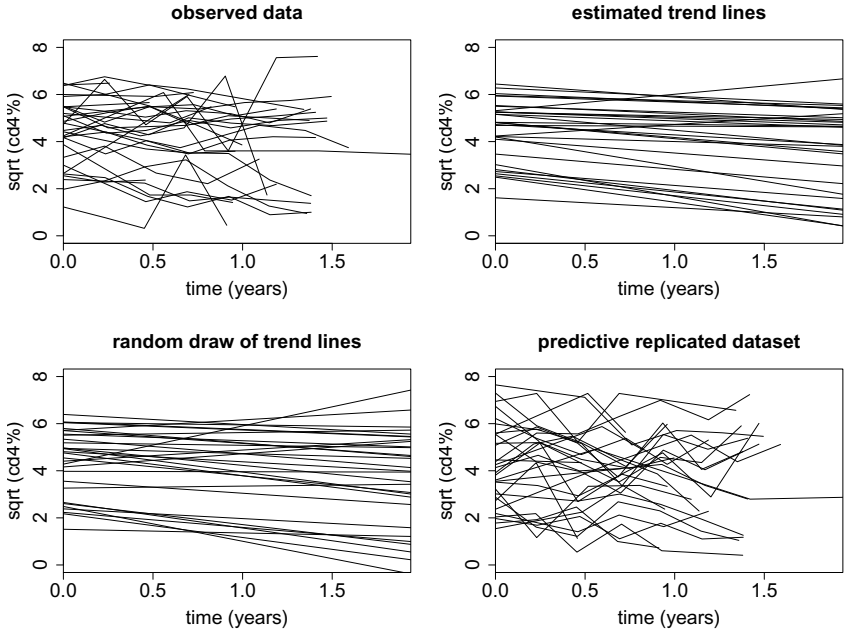

Figure 20.5 *(a) Progression of CD4 percentage over time (on the square root scale) for 83 untreated children j in the HIV study; (b) individual trend lines $\hat{\alpha}_j + \hat{\beta}_j t$ (posterior mean estimates from multilevel model); (c) a single posterior draw from the set of individual trend lines $\alpha_j + \beta_j t$; (d) a replicated dataset $(\tilde{y}_{jt})$ simulated from the posterior predictive distribution.*

at least 80% that the average estimated treatment effect is statistically significant at the 95% level.

A hypothesized model of treatment effects. To set up this power calculation we need to make assumptions about the true treatment effect and also specify all the other parameters that characterize the study. Our analysis of the HIV-positive children who did not receive zinc found an average decline in CD4 (on the square root scale) of 0.5 per year. We shall suppose in our power calculation that the true effect of the treatment is to reduce this average decline to zero.

We now set up a model for the hypothetical treatment and control data. So far, we have fitted a model to "controls," but that model can be used to motivate hypotheses for effects of treatments applied after the initial measurement ($t = 0$). To start with, the parameters α_j, β_j cleanly separate into an intercept that is unaffected by the treatment (and can thus be interpreted as an unobserved unit-level characteristic) and a slope β_j that is potentially affected. A model of linear trends can then be written as

$$\begin{aligned} y_i &\sim \text{N}(\alpha_{j[i]} + \beta_{j[i]} t_i,\, \sigma_y^2), \text{ for } i = 1, \ldots, n \\ \begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} &\sim \text{N}\left(\begin{pmatrix} \gamma_0^\alpha \\ \gamma_0^\beta + \gamma_1^\beta z_j \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix} \right), \text{ for } j = 1, \ldots, J, \end{aligned}$$

where

$$z_j = \begin{cases} 1 & \text{if child } i \text{ received the treatment} \\ 0 & \text{otherwise.} \end{cases}$$

The treatment z_j affects the slope β_j but not the intercept α_j because the treatment can have no effect at time zero. As noted, we shall suppose γ_0^β, the slope for controls,

to be -0.5, with a treatment effect of $\gamma_1^\beta = 0.5$. We complete the model by setting the other parameters to their estimated values from the control data: $\mu_\alpha = 4.8$, $\sigma_\alpha = 1.3$, $\sigma_y = 0.7$, $\sigma_\beta = 0.7$. For simplicity, we shall set ρ, the correlation between intercepts and slopes, to zero, although it was estimated at 0.1 from the actual data.

Design of the study. The next step in the power analysis is to specify the design of the study. We shall assume that J HIV-positive children will be randomly assigned into two treatments, with $J/2$ receiving regular care and $J/2$ receiving zinc supplements as well. We further assume that the children's CD4 percentages are measured every two months over a year (that is, seven measurements per child). We will now determine the J required for 80% power, if the true treatment effect is 0.5, as assumed above.

Quick power calculation for classical regression

We first consider a classical analysis, in which a separate linear regression is fitted for each child: $y_{jt} = \alpha_j + \beta_j t + \text{error}$. The trend estimates $\hat{\beta}_j$ would then be averaged for the children in the control and treatment groups, with the difference between the group mean trends being an estimated treatment effect. For simplicity, we assume the model is fitted separately for each child—that is, simple least squares, not a multilevel model.

This problem then has the structure of a simple classical sample size calculation, with the least squares estimate $\hat{\beta}_j$ being the single "data point" for each child j and an assumed effect size $\Delta = 0.5$. We must merely estimate σ, the standard deviation of the $\hat{\beta}_j$'s within each group, and we can determine the required total sample size as $J = (2 \cdot 2.8\sigma/\Delta)^2$.

If $\hat{\beta}_j$ were a perfect estimate of the child's trend parameter, then σ would simply be the standard deviation of the β_j's, or 0.7 from the assumptions we have made. However, we must also add the variance of estimation, which in this case (from the formula for least squares estimation with a single predictor) is $\frac{1}{\sqrt{(-3/6)^2+(-2/6)^2+\cdots+(3/6)^2}}\sigma_y = 1.13\sigma_y = 0.8$ (based on the estimate of $\sigma_y = 0.7$ from our multilevel model earlier). The total standard deviation of $\hat{\beta}_j$ is then $\sqrt{\sigma_\beta^2 + 1.13^2\sigma_y^2} = \sqrt{0.7^2 + 0.8^2} = 1.1$. The sample size required for 80% power to find a statistically significant difference in trends between the two groups is then $J = (2 \cdot 2.8 \cdot 1.1/0.5)^2 = 150$ children total (that is, 75 per group).

This sample size calculation is based on the assumption that the treatment would, on average, eliminate the observed decline in CD4 percentage. If instead we were to hypothesize that the treatment would cut the decline in half, the required sample size would quadruple, to a total of 600 children.

Power calculation for multilevel estimate using fake-data simulation

Power calculations for any model can be performed by simulation. This involves repeatedly simulating data from the hypothetical distribution that we expect our sampled data to come from (once we perform the intended study) and then fitting a multilevel model to each dataset. This can be computer-intensive, and practical compromises are sometimes needed so that the simulation can be performed in a reasonable time. Full simulation using Bugs is slow because it involves nested loops (100 or 1000 sets of fake data; for each, the looping of a Gibbs sampler required to

fit a model in Bugs). Instead, we fit the model to each fake dataset quickly using `lmer()`. We illustrate with the zinc treatment example.

Simulating the hypothetical data. The first step is to write a function in R that will generate data from the distribution assumed for the control children (based on our empirical evidence) and the distribution for the treated children (based on our assumptions about how their change in CD4 count might be different were they treated). This function generates data from a sample of J children (half treated, half controls), each measured K times during a 1-year period.

R code

```
CD4.fake <- function (J, K){
  time <- rep (seq(0,1,length=K), J)   # K measurements during the year
  person <- rep (1:J, each=K)          # person ID's
  treatment <- sample (rep (0:1, J/2))
  treatment1 <- treatment[person]
#                                      # hyperparameters:
  mu.a.true <- 4.8                     #   more generally, these could
  g.0.true <-  -.5                     #   be specified as additional
  g.1.true <-   .5                     #   arguments to the function
  sigma.y.true <-  .7
  sigma.a.true <- 1.3
  sigma.b.true <-  .7
#                                      # person-level parameters
  a.true <- rnorm (J, mu.a.true, sigma.a.true)
  b.true <- rnorm (J, g.0.true + g.1.true*treatment, sigma.b.true)
#                                      # data
  y <- rnorm (J*K, a.true[person] + b.true[person]*time, sigma.y.true)
  return (data.frame (y, time, person, treatment1))
}
```

The function returns a data frame with the simulated measurements along with the input variables needed to fit a model to the data and estimate the average treatment effect, γ_1. We save treatment as a data-level predictor (which we call `treatment1`) because this is how it must be entered into `lmer()`.

Fitting the model and checking the power. Next we can embed the fake-data simulation `CD4.fake()` in a loop to simulate 1000 sets of fake data; for each, we fit the model and obtain confidence intervals for the parameter of interest:

R code

```
CD4.power <- function (J, K, n.sims=1000){
  signif <- rep (NA, n.sims)
  for (s in 1:n.sims){
    fake <- CD4.fake (J, K)
    lme.power <- lmer (y ~ time + time:treatment1 +
      (1 + time | person), data=fake)
    theta.hat <- fixef(lme.power)["time:treatment1"]
    theta.se <- se.fixef(lme.power)["time:treatment1"]
    signif[s] <- (theta.hat - 2*theta.se) > 0     # returns TRUE or FALSE
  }
  power <- mean (signif)                          # proportion of TRUE
  return (power)
}
```

This function has several features that might need explaining:

- The function definition sets the number of simulations to the default value of 1000. So if `CD4.power()` is called without specifying the `n.sims` argument, it will automatically run 1000 simulations.

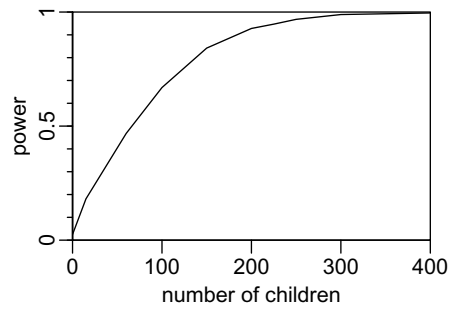

Figure 20.6 *Power (that is, the probability that estimated treatment effect is statistically significantly positive) as a function of number of children, J, for the hypothetical zinc study, as computed using fake-data simulation with multilevel inference performed by* `lmer()`. *The simulations are based on particular assumptions about the treatment effect and the variation among children and among measurements within children. We also have assumed $K = 7$ measurements for each child during the year of the study, a constraint determined by the practicalities of the experiment. Reading off the curve, 80% power is achieved at approximately $J = 130$.*

- The `lmer()` call includes the interaction `time:treatment1` and the main effect `time` but *not* the main effect `treatment1`. This allows the treatment to affect the slope but not the intercept, which is appropriate since the treatment is performed after time 0.
- The data frame `fake` is specified as an argument to `lmer()` so that the analysis knows what dataset to use.
- We assume the estimated treatment effect of the hypothetical study is statistically significantly positive if the lower bound of its 95% interval exceeds zero.
- The function returns the proportion of the 1000 simulations where the result is statistically significant; thus, the power (as computed via simulation) for a study with J children measured at K equally spaced times during the year.

Putting it all together to compute power as a function of sample size. Finally, we put the above simulation in a loop and compute the power at several different values of J, running from 20 to 400, and plot a curve displaying power as a function of sample size; the result is shown in Figure 20.6. Our quick estimate based on classical regression was that 80% power is achieved with $J = 150$ children (75 in each treatment group) also applies to the multilevel model in this case. The classical computation works in this case because the treatment is at the group level (in this example, persons are the groups, and CD4 measurements are the units) and the planned study is balanced.

At the two extremes:

- The power is 0.025 in the limit $J \to 0$. With a small enough sample, the treatment effect estimate is essentially random, and so there is a 2.5% chance that it is more than 2 standard errors above zero.
- Under the assumption that the true effect is positive, the power is 1 in the limit $J \to \infty$, at which point there are enough data to estimate the treatment effect perfectly.

Using simulation for power analyses allows for greater flexibility in study design. For instance, besides simply calculating how power changes as sample size increases, we might also have investigated a different kind of change in study design such as