a standard error of 0.35%, the estimate from the regression would then be statistically significant only if $\hat{\beta} > 0.7\%$ (or, strictly speaking, if $\hat{\beta} < -0.7\%$, but that latter possibility is highly unlikely given our assumptions). If the true coefficient is β, we would expect the estimate from the regression to possibly take on values in the range $\beta \pm 0.35\%$ (that is what is meant by "a standard error of 0.35%"), and thus if β truly equals 0.7%, we would expect $\hat{\beta}$ to exceed 0.7%, and thus achieve statistical significance, with a probability of 1/2—that is, 50% power. To get 80% power, we need the true β to be 2.8 standard errors from zero, so that there is an 80% probability that $\hat{\beta}$ is at least 2 standard errors from zero. If $\beta = 0.7\%$, then its standard error would have to be no greater than $0.7\%/2.8 = 0.25\%$, so that the survey would need a sample size of $(1.9\%/0.25\%)^2 \cdot 1192 = 70{,}000$.

This power calculation is only provisional, however, because it makes the very strong assumption that the β is equal to 0.7%, the estimate that we happened to obtain from our survey. But the estimate from the regression is $0.7\% \pm 1.9\%$, which implies that these data are consistent with a low, zero, or even negative value of the true β (or, in the other direction, a true value that is greater than the point estimate of 0.7%). If the true β is actually less than 0.7%, then even a sample size of 70,000 will be insufficient for 80% power.

This is not to say the power calculation is useless but just to point out that, even when done correctly, it is based on an assumption that is inherently untestable from the available data (hence the need for a larger study). So we should not necessarily expect statistical significance from a proposed study, even if the sample size has been calculated correctly.

20.4 Multilevel power calculation for cluster sampling

With multilevel data structures and models, power calculations become more complicated because there is the option to set the sample size at each level. In a cluster sampling design, one can choose the number of clusters to sample and the number of units to sample within each cluster. In a longitudinal study, one can choose the number of persons to study and the frequency of measurement of each person. Options become even more involved for more complicated designs, such as those involving treatments at different levels. We illustrate here with examples of quick calculations for a survey and an experiment and then in Section 20.5 discuss a general approach for power calculations using simulations.

Standard deviation of the mean of clustered data

Consider a survey in which it is desired to estimate the average value of y in some population, and data are collected from J equally sized clusters selected at random from a larger population, with m units measured from each sampled cluster, so that the total sample size is $n = Jm$.[1] In this symmetric design, the estimate for the population total is simply the sample mean, $\bar{y}$. If the number of clusters in the population is large compared to J, and the number of units within each cluster is large compared to m, then the standard error of $\bar{y}$ is

$$\text{standard error of } \bar{y} = \sqrt{\sigma_y^2/n + \sigma_\alpha^2/J}. \tag{20.1}$$

[1] In the usual notation for survey sampling, one might use a and A for the number of clusters in the sample and population, respectively. Here we use the capital letter J to indicate the number of selected clusters to be consistent with our general multilevel-modeling notation of J for the number of groups in the data.

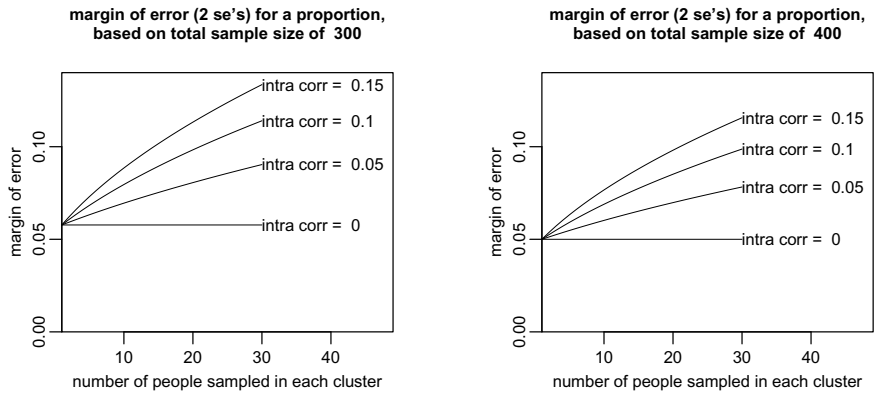

Figure 20.4 *Margin of error for inferences for a proportion as estimated from a cluster sample, as a function of cluster size and intraclass correlation, for two different proposed values of total sample size. The lines on the graphs do* not *represent a fitted model; they are based on analytical calculations using the variance formulas for cluster sampling.*

(The separate variance parameters σ_y^2 and σ_α^2, needed for the power calculations, can be estimated from the cluster-sampled data using a multilevel model.)

This formula can also be rewritten as

$$\text{standard error of } \bar{y} = \sqrt{\frac{\sigma_{\text{total}}^2}{Jm}[1 + (m-1)\text{ICC}]}, \tag{20.2}$$

where σ_{total} represents the standard deviation of all the data (mixing all the groups; thus $\sigma_{\text{total}}^2 = \sigma_y^2 + \sigma_\alpha^2$ for this simple model), and ICC is the *intraclass correlation*,

$$\text{intraclass correlation: } \text{ICC} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_y^2}, \tag{20.3}$$

the fraction of total variation in the data that is accounted for by between-group variation. The intraclass correlation can also be thought of as the correlation among units within the same group. Formulas (20.1) and (20.2) provide some intuition regarding the extent to which clustering can affect our standard errors. The greater the correlation among units within a group (that is, the bigger ICC is) the greater the impact on the standard error. If there is no intraclass correlation (that is, ICC = 0) the standard error of $\bar{y}$ is simply $\sigma_{\text{total}}/\sqrt{n}$.

Example of a sample size calculation for cluster sampling

We illustrate sample size calculations for cluster sampling with a design for a proposed study of residents of New York City. The investigators were planning to study approximately 300 or 400 persons sampled for convenience from 10 or 20 U.S. Census tracts, and they wanted to get a sense of how much error the clustering was introducing into the estimation. The number of census tracts in the city and the population of each tract are large enough that (20.1) was a reasonable approximation.

Figure 20.4 shows the margin of error for $\bar{y}$ from this formula, as a function of the sample size within clusters, for several values of the intraclass correlation. When the correlation is zero, the clustering is irrelevant and the margin of error only depends on the total sample size, n. For positive values of intraclass correlation (so that people within a census tract are somewhat similar to each other, on average),