Chapter 16

Design and sample size decisions

This chapter is a departure from the rest of the book, which focuses on data analysis: building, fitting, understanding, and evaluating models fit to existing data. In the present chapter, we consider the design of studies, in particular asking the question of what sample size is required to estimate a quantity of interest to some desired precision. We focus on the paradigmatic inferential tasks of estimating population averages, proportions, and comparisons in sample surveys, or estimating treatment effects in experiments and observational studies. However, the general principles apply for other inferential goals such as prediction and data reduction. We present the relevant algebra and formulas for sample size decisions and demonstrating with a range of examples, but we also criticize the standard design framework of "statistical power," which when studied naively yields unrealistic expectations of success and can lead to the design of ineffective, noisy studies. As we frame it, the goal of design is not to attain statistical significance with some high probability, but rather to have a sense—before and after data have been collected—about what can realistically be learned from statistical analysis of an empirical study.

16.1 The problem with statistical power

Statistical *power* is defined as the probability, before a study is performed, that a particular comparison will achieve "statistical significance" at some predetermined level (typically a p-value below 0.05), given some assumed true effect size. A power analysis is performed by first hypothesizing an effect size, then making some assumptions about the variation in the data and the sample size of the study to be conducted, and finally using probability calculations to determine the chance of the p-value being below the threshold.

The conventional view is that you should avoid low-power studies because they are unlikely to succeed. This, for example, comes from an influential paper in criminology:

> Statistical power provides the most direct measure of whether a study has been designed to allow a fair test of its research hypothesis. When a study is underpowered it is unlikely to yield a statistically significant result even when a relatively large program or intervention effect is found.

This statement is correct but too simply presents statistical significance as a goal.

To see the problem with aiming for statistical significance, suppose that a study is low power but can be performed for free, or for a cost that it is very low compared to the potential benefits that would arise from a research success. Then a researcher might conclude that a lower-power study is still worth doing, that it is a gamble worth undertaking.

The traditional power threshold is 80%; funding agencies are reluctant to approve studies that are not deemed to have at least an 80% chance of obtaining a statistically significant result. But under a simple cost-benefit calculation, there would be cases where 50% power, or even 10% power, would suffice, for simple studies such as psychology experiments where human and dollar costs are low. Hence, when costs are low, researchers are often inclined to roll the dice, on the belief that a successful finding could potentially bring large benefits (to society as well as to the researcher's career). But this is not necessarily a good idea, as we discuss next.

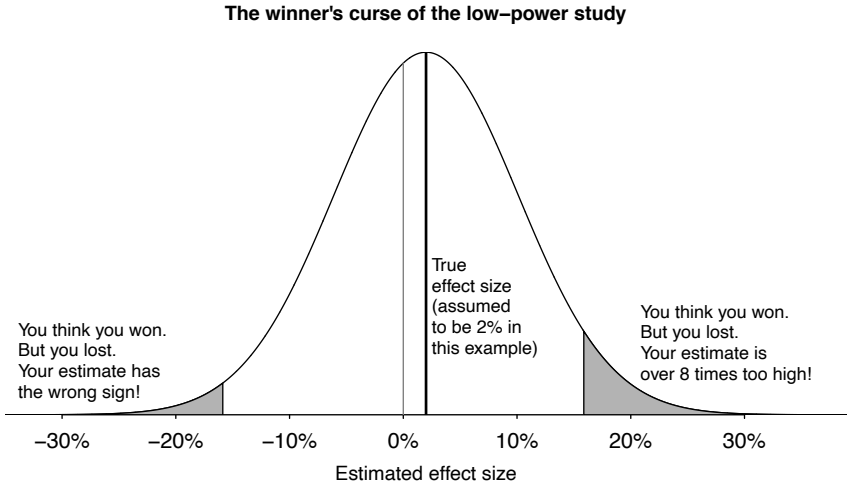

Figure 16.1 *When the effect size is small compared to the standard error, statistical power is low. In this diagram, the bell-shaped curve represents the distribution of possible estimates, and the gray shaded zones correspond to estimates that are "statistically significant" (at least two standard errors away from zero). In this example, statistical significance is unlikely to be achieved, but in the rare cases where it does happen, it is highly misleading: there is a large chance the estimate has the wrong sign (a type S error) and, in any case, the magnitude of the effect size will be vastly overstated (a type M error) if it happens to be statistically significant. Thus, what would naively appear to be a "win" or a lucky draw—a statistically significant result from a low-power study—is, in the larger sense, a loss to science and to policy evaluation.*

The winner's curse in low-power studies

The problem with the conventional reasoning is that, in a low-power study, the seeming "win" of statistical significance can actually be a trap. Economists speak of a "winner's curse" in which the highest bidder in an auction will, on average, be overpaying. Research studies—even randomized experiments—suffer from a similar winner's curse, that by focusing on comparisons that are statistically significant, we (the scholarly community as well as individual researchers) get a systematically biased and over-optimistic picture of the world.

Put simply, when signal is low and noise is high, statistically significant patterns in data are likely to be wrong, in the sense that the results are unlikely to replicate.

To put it in technical terms, statistically significant results are subject to type M and type S errors, as described in Section 4.4. Figure 16.1 illustrates for a study where the true effect could not realistically be more than 2 percentage points and is estimated with a standard error of 8.1 percentage points. We can examine the statistical properties of the estimate using the normal distribution: conditional on it being statistically significant (that is, at least two standard errors from zero), the estimate has at least a 24% probability of being in the wrong direction and is, by necessity, over 8 times larger than the true effect.

A study with these characteristics has essentially no chance of providing useful information, and we can say this even before the data have been collected. Given the numbers above for standard error and possible effect size, the study has a power of at most 6% (see Exercise 16.4), but it would be misleading to say it has even a 6% chance of success. From the perspective of scientific learning, the real failures are the 6% of the time that the study appears to succeed, in that these correspond to ridiculous overestimates of treatment effects that are likely to be in the wrong direction as well. In such an experiment, to win is to lose.

Thus, a key risk for a low-power study is not so much that it has a small chance of succeeding, but rather that an apparent success merely masks a larger failure. Publication of noisy findings in

turn can contribute to the replication crisis when these fragile claims collapse under more careful analysis or do not show up in attempted replications, as discussed in Section 4.5.

Hypothesizing an effect size

The other challenge is that any power analysis or sample size calculations is conditional on an assumed effect size, and this is something that is the target of the study and is thus never known ahead of time.

There are different ways to choose an effect size for performing an analysis of a planned study design. One strategy, which we demonstrate in Section 16.5, is to try a range of values consistent with the previous literature on the topic. Another approach is to decide what magnitude of effect would be of practical interest: for example, in a social intervention we might feel that we are only interested in pursuing a particular treatment if it increases some outcome by at least 10%; we could then perform a design analysis to see what sample size would be needed to reliably detect an effect of that size.

One common practice that we do *not* recommend is to make design decisions based on the estimate from a single noisy study. Section 16.3 gives an example of how one can use a patchwork of information from earlier studies to make informed judgments about statistical power and sample size.

16.2 General principles of design, as illustrated by estimates of proportions

Effect sizes and sample sizes

In designing a study, it is generally better, if possible, to double the effect size θ than to double the sample size n, since standard errors of estimation decrease with the square root of the sample size. This is one reason, for example, why potential toxins are tested on animals at many times their exposure levels in humans; see Exercise 16.8.

Studies are designed in several ways to maximize effect size:

- In drug studies, setting doses as low as ethically possible in the control group and as high as ethically possible in the experimental group.
- To the extent possible, choosing individuals that are likely to respond strongly to the treatment. For example, an educational intervention in schools might be performed on poorly performing classes in each grade, for which there will be more room for improvement.

In practice, this advice cannot be followed completely. Sometimes it can be difficult to find an intervention with *any* noticeable positive effect, let alone to design one where the effect would be doubled. Also, when treatments in an experiment are set to extreme values, generalizations to more realistic levels can be suspect. Further, treatment effects discovered on a sensitive subgroup may not generalize to the entire population. But, on the whole, conclusive effects on a subgroup are generally preferred to inconclusive but more generalizable results, and so conditions are usually set up to make effects as large as possible.

Published results tend to be overestimates

There are various reasons why we would typically expect future effects to be smaller than published estimates. First, as noted just above, interventions are often tested on people and in scenarios where they will be most effective—indeed, this is good design advice—and effects will be smaller in the general population "in the wild." Second, results are more likely to be reported and more likely to be published when they are "statistically significant," which leads to overestimation: type M errors, as discussed in Section 4.4. Some understanding of the big picture is helpful when considering how to interpret the results of published studies, even beyond the uncertainty captured in the standard error.

Design calculations

Before data are collected, it can be useful to estimate the precision of inferences that one expects to achieve with a given sample size, or to estimate the sample size required to attain a certain precision. This goal is typically set in one of two ways:

- Specifying the standard error of a parameter or quantity to be estimated, or
- specifying the probability that a particular estimate will be "statistically significant," which typically is equivalent to ensuring that its 95% confidence interval will exclude the null value.

In either case, the sample size calculation requires assumptions that typically cannot really be tested until the data have been collected. Sample size calculations are thus inherently hypothetical.

Sample size to achieve a specified standard error

To understand these two kinds of calculations, consider the simple example of estimating the proportion of the population who support the death penalty (under a particular question wording). Suppose we suspect the population proportion is around 60%. First, consider the goal of estimating the true proportion p to an accuracy (that is, standard error) of no worse than 0.05, or 5 percentage points, from a simple random sample of size n. The standard error of the mean is $\sqrt{p(1-p)/n}$. Substituting the guessed value of 0.6 for p yields a standard error of $\sqrt{0.6 * 0.4/n} = 0.49/\sqrt{n}$, and so we need $0.49/\sqrt{n} \leq 0.05$, or $n \geq 96$. More generally, we do not know p, so we would use a conservative standard error of $\sqrt{0.5 * 0.5/n} = 0.5/\sqrt{n}$, so that $0.5/\sqrt{n} \leq 0.05$, or $n \geq 100$.

Sample size to achieve a specified probability of obtaining statistical significance

Second, suppose we have the goal of demonstrating that more than half the population supports the death penalty—that is, that $p > 1/2$—based on the estimate $\hat{p} = y/n$ from a sample of size n. As above, we shall evaluate this under the hypothesis that the true proportion is $p = 0.60$, using the conservative standard error for $\hat{p}$ of $\sqrt{0.5 * 0.5/n} = 0.5/\sqrt{n}$. The 95% confidence interval for p is $[\hat{p} \pm 1.96 * 0.5/\sqrt{n}]$, and classically we would say we have demonstrated that $p > 1/2$ if the interval lies entirely above 1/2; that is, if $\hat{p} > 0.5 + 1.96 * 0.5/\sqrt{n}$. The estimate must be at least 1.96 standard errors away from the comparison point of 0.5.

A simple, but not quite correct, calculation, would set $\hat{p}$ to the hypothesized value of 0.6, so that the requirement is $0.6 > 0.5 + 1.96 * 0.5/\sqrt{n}$, or $n > (1.96 * 0.5/0.1)^2 = 96$. This is mistaken, however, because it confuses the assumption that $p = 0.6$ with the claim that $\hat{p} > 0.6$. In fact, if $p = 0.6$, then $\hat{p}$ depends on the sample, and it has an approximate normal distribution with mean 0.6 and standard deviation $\sqrt{0.6 * 0.4/n} = 0.49/\sqrt{n}$; see the top half of Figure 16.2.

To determine the appropriate sample size, we must specify the desired *power*—that is, the probability that a 95% interval will be entirely above the comparison point of 0.5. Under the assumption that $p = 0.6$, choosing $n = 96$ yields 50% power: there is a 50% chance that $\hat{p}$ will be more than 1.96 standard deviations away from 0.5, and thus a 50% chance that the 95% interval will be entirely greater than 0.5.

The conventional level of power in sample size calculations is 80%: the goal is to choose n such that 80% of the possible 95% confidence intervals will not include 0.5. When n is increased, the estimate becomes closer (on average) to the true value, and the width of the confidence interval decreases. Both these effects (decreasing variability of the estimator and narrowing of the confidence interval) can be seen in going from the top half to the bottom half of Figure 16.2.

To find the value of n such that exactly 80% of the estimates will be at least 1.96 standard errors from 0.5, we need

$$0.5 + 1.96 * \text{s.e.} = 0.6 - 0.84 * \text{s.e.}$$

Some algebra then yields $(1.96 + 0.84) * \text{s.e.} = 0.1$. We can then substitute s.e. $= 0.5/\sqrt{n}$ and solve for n, as we discuss next.

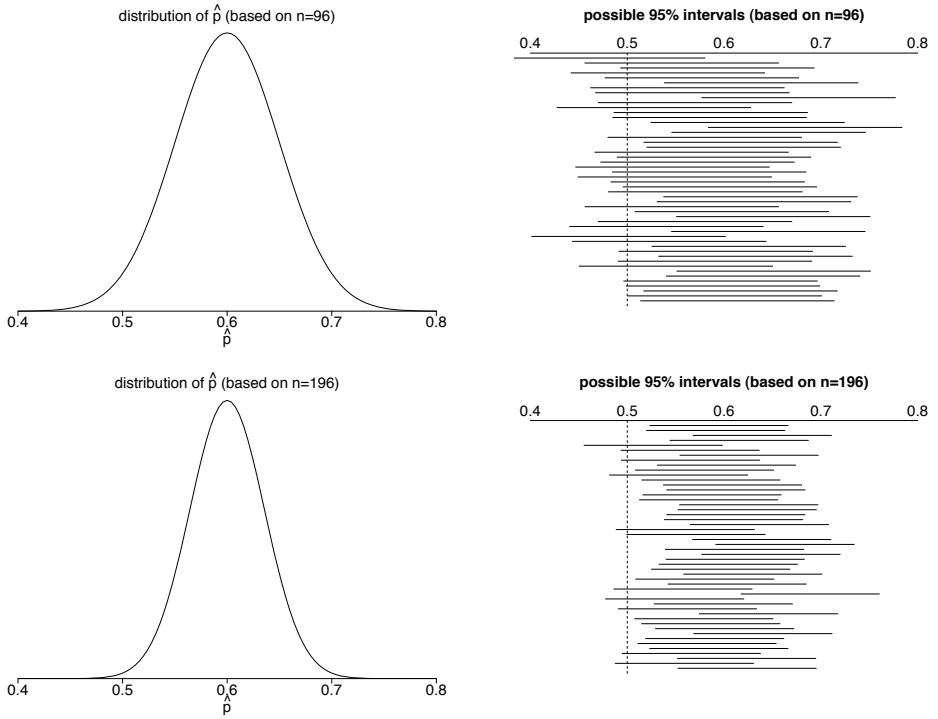

Figure 16.2 *Illustration of simple sample size calculations.* Top row: *(left) distribution of the sample proportion $\hat{p}$ if the true population proportion is $p = 0.6$, based on a sample size of 96; (right) several possible 95% intervals for p based on a sample size of 96. The power is 50%—that is, the probability is 50% that a randomly generated interval will be entirely to the right of the comparison point of 0.5.* Bottom row: *corresponding graphs for a sample size of 196. Here the power is 80%.*

In summary, *to have 80% power, the true value of the parameter must be 2.8 standard errors away from the comparison point*: the value 2.8 is 1.96 from the 95% interval, plus 0.84 to reach the 80$^{\text{th}}$ percentile of the normal distribution. The bottom row of Figures 16.2 and 16.3 illustrate: with $n = (2.8 * 0.49/0.1)^2 = 196$, and if the true population proportion is $p = 0.6$, there is an 80% chance that the 95% confidence interval will be entirely greater than 0.5, thus conclusively demonstrating that more than half the people support the death penalty.

These calculations are only as good as their assumptions; in particular, one would generally not know the true value of p before doing the study. Nonetheless, design analyses can be useful in giving a sense of the size of effects that one could reasonably expect to demonstrate with a study of given size. For example, a survey of size 196 has 80% power to demonstrate that $p > 0.5$ if the true value is 0.6, and it would easily detect the difference if the true value were 0.7; but if the true p were equal to 0.56, say, then the difference would be only $0.06/(0.5/\sqrt{196}) = 1.6$ standard errors away from zero, and it would be likely that the 95% interval for p would include 0.5, even in the presence of this true effect. Thus, if the goal of the survey is to conclusively detect a difference from 0.5, it would probably not be wise to use a sample of only $n = 196$ unless we suspect the true p is at least 0.6. Such a small survey would not have the power to reliably detect differences of less than 0.1.

Estimates of hypothesized proportions

The standard error of a proportion p, if it is estimated from a simple random sample of size n, is $\sqrt{p(1-p)/n}$, which has an upper bound of $0.5/\sqrt{n}$. This upper bound is very close to the actual standard error for a wide range of probabilities p near 1/2: for example, if the probability is 0.5, then the standard error is $\sqrt{0.5 * 0.5/n} = 0.5/\sqrt{n}$ exactly; if probabilities are 60/40, then we get

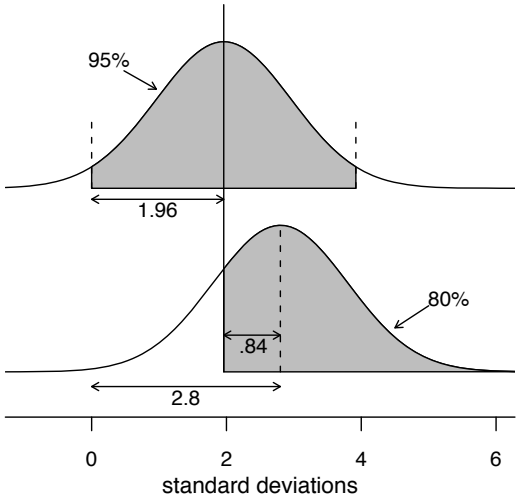

Figure 16.3 *Sketch illustrating that, to obtain 80% power for a 95% confidence interval, the true effect size must be at least 2.8 standard errors from zero (assuming a normal distribution for estimation error). The top curve shows that the estimate must be at least 1.96 standard errors from zero for the 95% interval to be entirely positive. The bottom curve shows the distribution of the parameter estimates that might occur, if the true effect size is 2.8. Under this assumption, there is an 80% probability that the estimate will exceed 1.96. The two curves together show that the lower curve must be centered all the way at 2.8 to get an 80% probability that the 95% interval will be entirely positive.*

$\sqrt{0.6 * 0.4/n} = 0.49/\sqrt{n}$; and if probabilities are 70/30, then we get $\sqrt{0.7 * 0.3/n} = 0.46/\sqrt{n}$, which is still not far from $0.5/\sqrt{n}$.

If the goal is a specified standard error, then the required sample size is determined conservatively by s.e. $= 0.5/\sqrt{n}$, so that $n = (0.5/\text{s.e.})^2$ or, more precisely, $n = p(1-p)/(\text{s.e.})^2$. If the goal is 80% power to distinguish p from a specified value p_0, then a conservative required sample size is that needed for the true parameter value to be 2.8 standard errors from zero; solving for this standard error yields $n = (2.8 * 0.5/(p - p_0))^2$ or, more precisely, $n = p(1-p)(2.8/(p-p_0))^2$.

Simple comparisons of proportions: equal sample sizes

The standard error of a difference between two proportions is, by a simple probability calculation, $\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}$, which has an upper bound of $0.5\sqrt{1/n_1 + 1/n_2}$. If we assume $n_1 = n_2 = n/2$ (equal sample sizes in the two groups), the upper bound on the standard error becomes simply $1/\sqrt{n}$. A specified standard error can then be attained with a sample size of $n = 1/(\text{s.e.})^2$. If the goal is 80% power to distinguish between hypothesized proportions p_1 and p_2 with a study of size n, equally divided between the two groups, a conservative sample size is $n = ((2.8/(p_1 - p_2))^2$ or, more precisely, $n = 2(p_1(1-p_1) + p_2(1-p_2))(2.8/(p_1-p_2))^2$.

For example, suppose we suspect that the death penalty is 10% more popular in the United States than in Canada, and we plan to conduct surveys in both countries on the topic. If the surveys are of equal sample size, $n/2$, how large must n be so that there is an 80% chance of achieving statistical significance, if the true difference in proportions is 10%? The standard error of $\hat{p}_1 - \hat{p}_2$ is approximately $1/\sqrt{n}$, so for 10% to be 2.8 standard errors from zero, we must have $n > (2.8/0.10)^2 = 784$, or a survey of 392 people in each country.

Simple comparisons of proportions: unequal sample sizes

In epidemiology, it is common to have unequal sample sizes in comparison groups. For example, consider a study in which 20% of units are exposed and 80% are controls.

First, consider the goal of estimating the difference between the exposed and control groups, to some specified precision. The standard error of the difference is $\sqrt{p_1(1-p_1)/(0.2n) + p_2(1-p_2)/(0.8n)}$, and this expression has an upper bound of $0.5\sqrt{1/(0.2n) + 1/(0.8n)} = 0.5\sqrt{1/(0.2) + 1/(0.8)}/\sqrt{n} = 1.25/\sqrt{n}$. A specified standard error can then be attained with a sample size of $n = (1.25/\text{s.e.})^2$.

Second, suppose we want sufficient total sample size n to achieve 80% power to detect a difference of 10%, again with 20% of the sample size in one group and 80% in the other. Again, the standard error of $\hat{p}_1 - \hat{p}_2$ is bounded by $1.25/\sqrt{n}$, so for 10% to be 2.8 standard errors from zero, we must have $n > (2.8 * 1.25/0.10)^2 = 1225$, or 245 cases and 980 controls.

16.3 Sample size and design calculations for continuous outcomes

Example: Zinc experiments

Sample size calculations proceed much the same way with continuous outcomes, with the added difficulty that the population standard deviation must also be specified along with the hypothesized effect size. We shall illustrate with a proposed experiment adding zinc to the diet of HIV-positive children in South Africa. In various other populations, zinc and other micronutrients have been found to reduce the occurrence of diarrhea, which is associated with immune system problems, as well as to slow the progress of HIV. We first consider the one-sample problem—how large a sample size we would expect to need to measure various outcomes to a specified precision—and then move to two-sample problems comparing treatment to control groups.

Estimates of means

Suppose we are trying to estimate a population mean value θ from data $y_1, \ldots, y_n$, a random sample of size n. The quick estimate of θ is the sample mean, $\bar{y}$, which has a standard error of $\sigma/\sqrt{n}$, where σ is the standard deviation of y in the population. So if the goal is to achieve a specified s.e. for $\bar{y}$, then the sample size must be at least $n = (\sigma/\text{s.e.})^2$. If the goal is 80% power to distinguish θ from a specified value θ_0, then a conservative required sample size is $n = (2.8\,\sigma/(\theta - \theta_0))^2$.

The t distribution and uncertainty in standard deviations

In this section, we perform all design analyses using the normal distribution, which is appropriate for linear regression when the residual standard deviation σ is known. For very small studies, though, degrees of freedom are low, the residual standard deviation is not estimated precisely from data, and inferential uncertainties (confidence intervals or posterior intervals) follow the t distribution. In that case, the value 2.8 needs to be replaced with a larger number to capture this additional source of uncertainty. For example, when designing a study comparing two groups of 6 patients each, the degrees of freedom are 10 (calculated as 12 data points minus two coefficients being estimated; see the beginning of Section 4.4), and the normal distributions in the power calculations are replaced by t_{10}. In R, `qnorm(0.8) + qnorm(0.975)` yields the value 2.8, while `qt(0.8,10) + qt(0.975,10)` yields the value 3.1, so we would replace 2.8 by 3.1 in the calculations for 80% power. We usually don't worry about the t correction because it is minor except when sample sizes are very small.

Simple comparisons of means

The standard error of $\bar{y}_1 - \bar{y}_2$ is $\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$. If we make the restriction $n_1 = n_2 = n/2$ (equal sample sizes in the two groups), the standard error becomes simply s.e. $= \sqrt{2(\sigma_1^2 + \sigma_2^2)/n}$. A specified standard error can then be attained with a sample size of $n = 2(\sigma_1^2 + \sigma_2^2)/(\text{s.e.})^2$. If we further suppose that the variation is the same within each of the groups ($\sigma_1 = \sigma_2 = \sigma$), then s.e. $= 2\sigma/\sqrt{n}$, and the required sample size is $n = (2\sigma/\text{s.e.})^2$.

If the goal is 80% power to detect a difference of Δ, with a study of size n, equally divided

Rosado et al. (1997), Mexico

Treatment	Sample size	Avg. # episodes in a year ± s.e.
placebo	56	1.1 ± 0.2
iron	54	1.4 ± 0.2
zinc	54	0.7 ± 0.1
zinc + iron	55	0.8 ± 0.1

Ruel et al. (1997), Guatemala

Treatment	Sample size	Avg. # episodes per 100 days [95% c.i.]
placebo	44	8.1 [5.8, 10.2]
zinc	45	6.3 [4.2, 8.9]

Lira et al. (1998), Brazil

Treatment	Sample size	% days with diarrhea	Prevalence ratio [95% c.i.]
placebo	66	5%	1
1 mg zinc	68	5%	1.00 [0.72, 1.40]
5 mg zinc	71	3%	0.68 [0.49, 0.95]

Muller et al. (2001), West Africa

Treatment	Sample size	# days with diarrhea/ total # days
placebo	329	997/49 021 = 0.020
zinc	332	869/49 086 = 0.018

Figure 16.4 *Results from various experiments studying zinc supplements for children with diarrhea. We use this information to hypothesize the effect size Δ and within-group standard deviation σ for our planned experiment.*

between the two groups, then the required sample size is $n = 2(\sigma_1^2 + \sigma_2^2)(2.8/\Delta)^2$. If $\sigma_1 = \sigma_2 = \sigma$, this simplifies to $(5.6\,\sigma/\Delta)^2$.

For example, consider the effect of zinc supplements on young children's growth. Results of published studies suggest that zinc can improve growth by approximately 0.5 standard deviations. That is, $\Delta = 0.5\sigma$ in the our notation. To have 80% power to detect an effect size, it would be sufficient to have a total sample size of $n = (5.6/0.5)^2 = 126$, or $n/2 = 63$ in each group.

Estimating standard deviations using results from previous studies

Sample size calculations for continuous outcomes are based on estimated effect sizes and standard deviations in the population—that is, Δ and σ. Guesses for these parameters can be estimated or deduced from previous studies. We illustrate with the design of a study to estimate the effects of zinc on diarrhea in children. Various experiments have been performed on this topic—Figure 16.4 summarizes the results, which we shall use to get a sense of the sample size required for our study.

We consider the studies reported in Figure 16.4 in order. For Rosado et al. (1997), we estimate the effect of zinc by averaging over the iron and no-iron cases, thus an estimated Δ of $\frac{1}{2}(1.1+1.4) - \frac{1}{2}(0.7+0.8) = 0.5$ episodes in a year, with a standard error of $\sqrt{\frac{1}{4}(0.2^2 + 0.2^2) + \frac{1}{4}(0.1^2 + 0.1^2)} = 0.15$. From this study, we estimate that zinc reduces diarrhea in that population by an average of about 0.3 to 0.7 episodes per year. Next, we deduce the within-group standard deviations σ using the formula s.e.$= \sigma/\sqrt{n}$; thus the standard deviations are $0.2 * \sqrt{56} = 1.5$ for the placebo group and are for 1.5, 0.7, and 0.7 for the other three groups. The number of episodes is bounded below by zero, so it makes sense that when the mean level goes down, the standard deviation decreases also.

Assuming an effect size of $\Delta = 0.5$ episodes per year and within-group standard deviations of 1.5 and 0.7 for the control and treatment groups, we can evaluate the power of a future study with $n/2$ children in each group. The estimated difference would have a standard error of $\sqrt{1.5^2/(n/2) + 0.7^2/(n/2)} = 2.4/\sqrt{n}$, and so for the effect size to be at least 2.8 standard errors away from zero (and thus to have 80% power to attain statistical significance), n would have to be at least $(2.8 * 2.4/0.5)^2 = 180$ people in the two groups.

Now turning to the Ruel et al. (1997) study, we first see that rates of diarrhea—for control and treated children both—are much higher than in the previous study: 8 episodes per hundred days, which corresponds to 30 episodes per year, more than 20 times the rate in the earlier group. We are dealing with very different populations here. In any case, we can divide the uncertainty interval widths by 4 to get standard errors—thus, 1.1 for the placebo group and 1.2 for the treated group—yielding an estimated treatment effect of 1.8 with standard error 1.6, which is consistent with a treatment effect of nearly zero or as high as about 4 episodes per 100 days. When compared to the average observed rate in the control group, the estimated treatment effect from this study is about half that of the Rosado et al. (1997) experiment: $1.8/8.1 = 0.22$, compared to $0.5/1.15 = 0.43$, which suggests a higher sample size might be required. However, the wide uncertainty bounds of the Ruel et al. (1997) study make it consistent with the larger effect size.

Next, Lira et al. (1998) report the average percentage of days with diarrhea of children in the control and two treatment groups corresponding to a low (1 mg) or high (5 mg) dose of zinc. We shall consider only the 5 mg condition, as this is closer to the treatment for our experiment. The estimated effect of the treatment is to multiply the number of days with diarrhea by 68%—that is, a reduction of 32%, which again is consistent with the approximate 40% decrease found in the first study. To make a power calculation, we first convert the uncertainty interval $[0.49, 0.95]$ for this multiplicative effect to the logarithmic scale—thus, an additive effect of $[-0.71, -0.05]$ on the logarithm—then divide by 4 to get an estimated standard error of 0.16 on this scale. The estimated effect of 0.68 is -0.38 on the log scale, thus 2.4 standard errors away from zero. For this effect size to be 2.8 standard errors from zero, we would need to increase the sample size by a factor of $(2.8/2.4)^2 = 1.4$, thus moving from approximately 70 children to approximately 100 in each of the two groups.

Finally, Muller et al. (2001) compare the proportion of days with diarrhea, which declined from 2.03% in the controls to 1.77% among children who received zinc. Unfortunately, no standard error is reported for this 13% decrease, and it is not possible to compute it from the information in the article. However, the estimates of within-group variation σ from the other studies would lead us to conclude that we would need a very large sample size to be likely to reach statistical significance, if the true effect size were only 10%. For example, from the Lira et al. (1998) study, we estimate a sample size of 100 in each group is needed to detect an effect of 32%; thus, to detect a true effect of 13%, we would need a sample size of $100 * (0.32/0.13)^2 = 600$.

These calculations are necessarily speculative; for example, to detect an effect of 10% (instead of 13%), the required sample size would be $100 * (0.32/0.10)^2 = 1000$ per group, a huge change considering the very small change in hypothesized treatment effects. Thus, it is misleading to think of these as required sample sizes. Rather, these calculations tell us how large the effects are that we could expect to have a good chance of discovering, given any specified sample size.

The first two studies in Figure 16.4 report the frequency of episodes, and the last two give the proportion of days with diarrhea, which is proportional to the frequency of episodes multiplied by the average duration of each episode. Other data (not shown here) show no effect of zinc on average duration, and so we treat all four studies as estimating the effects on frequency of episodes.

In conclusion, a sample size of about 100 per treatment group should give adequate power to detect an effect of zinc on diarrhea, if its true effect is to reduce the frequency, on average, by 30%–50% compared to no treatment. A sample size of 200 per group would have the same power to detect effects a factor $\sqrt{2}$ smaller, that is, effects in the 20%–35% range.

Including more regression predictors

Now suppose we are comparing treatment and control groups with additional pre-treatment data on the children (for example, age, height, weight, and health status at the start of the experiment). These can be included in a regression. For simplicity, consider a model with no interactions—that is, with coefficients for the treatment indicator and the other inputs—in which case, the treatment coefficient represents the causal effect, the comparison after adjusting for pre-treatment differences.

Sample size calculations for this new study are exactly as before, except that the within-group

standard deviation σ is replaced by the residual standard deviation of the regression. This can be hypothesized in its own right or in terms of the added predictive power of the pre-treatment data. For example, if we hypothesize a within-group standard deviation of 0.2, then a residual standard deviation of 0.14 would imply that half the variance within any group is explained by the regression model, which would actually be pretty good.

Adding relevant predictors should decrease the residual standard deviation and thus reduce the required sample size for any specified level of precision or power.

Estimation of regression coefficients more generally

More generally, sample sizes for regression coefficients and other estimands can be calculated using the rule that standard errors are proportional to $1/\sqrt{n}$; thus, if inferences exist under a current sample size, effect sizes can be estimated and standard errors extrapolated for other hypothetical samples.

We illustrate with the example of the survey earnings and height discussed in Chapter 4. The coefficient for the sex-earnings interaction in model (12.2) is plausible (a positive interaction, implying that an extra inch of height is worth 0.7% more for men than for women), but it is not statistically significant—the standard error is 1.9%, yielding a 95% interval of $[-3.1, 4.5]$, which contains zero.

How large a sample size would be needed for the coefficient on the interaction to be statistically significant? A simple calculation uses the fact that standard errors are proportional to $1/\sqrt{n}$. For a point estimate of 0.7% to achieve statistical significance, it would need a standard error of 0.35%, which would require the sample size to be increased by a factor of $(1.9\%/0.35\%)^2 = 29$. The original survey had a sample of 1192; this implies a required sample size of $29 * 1192 = 35\,000$.

To extend this to a power calculation, we suppose that the true β for the interaction is equal to 0.7% and that the standard error is as we have just calculated. With 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 $(1.9\%/0.25\%)^2 * 1192 = 70\,000$.

This design calculation is close to meaningless, however, because it makes the very strong assumption that the true value of β is 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 would be insufficient for 80% power.

This is not to say the design analysis 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. To put it another way, the value of the above calculations is *not* to tell us the power of the study that was just performed, or to choose a sample size of a new study, but rather to develop our intuitions of the relation between inferential uncertainty, standard error, and sample size.

Sample size, design, and interactions

Sample size is never large enough. As n increases, we can estimate more interactions, which typically are smaller and have relatively larger standard errors than main effects; for example, see the fitted regression on page 193 of log earnings on sex, standardized height, and their interaction. Estimating

interactions is similar to comparing coefficients estimated from subsets of the data (for example, the coefficient for height among men, compared to the coefficient among women), thus reducing power because the sample size for each subset is halved, and also the differences themselves may be small. As more data are included in an analysis, it becomes possible to estimate these interactions (or, using multilevel modeling, to include them and partially pool them as appropriate), so this is not a problem. We are just emphasizing that, just as you never have enough money, because perceived needs increase with resources, your inferential needs will increase with your sample size.

16.4 Interactions are harder to estimate than main effects

In causal inference, it is often important to study varying effects: for example, a treatment could be more effective for men than for women, or for healthy than for unhealthy patients. We are often interested in interactions in predictive models as well.

You need 4 times the sample size to estimate an interaction that is the same size as the main effect

Suppose a study is designed to have 80% power to detect a main effect at a 95% confidence level. As discussed earlier in this chapter, that implies that the true effect size is 2.8 standard errors from zero. That is, the z-score has a mean of 2.8 and standard deviation of 1, and there's an 80% chance that the z-score exceeds 1.96 (in R, `pnorm(2.8,1.96,1) = 0.8`).

Further suppose that an interaction of interest is the same size as the main effect. For example, if the average treatment effect on the entire population is θ, with an effect of $0.5\,\theta$ among women and $1.5\,\theta$ among men, then the interaction—the difference in treatment effect comparing men to women—is the same size as the main effect.

The standard error of an interaction is roughly *twice* the standard error of the main effect, as we can see from some simple algebra:

- The estimate of the main effect is $\bar{y}_T - \bar{y}_C$, and this has standard error $\sqrt{\sigma^2/(n/2) + \sigma^2/(n/2)} = 2\sigma/\sqrt{n}$; for simplicity we are assuming a constant variance within groups, which will typically be a good approximation for binary data, for example.
- The estimate of the interaction is $(\bar{y}_{T,\text{men}} - \bar{y}_{C,\text{men}}) - (\bar{y}_{T,\text{women}} - \bar{y}_{C,\text{women}})$, which has standard error $\sqrt{\sigma^2/(n/4) + \sigma^2/(n/4) + \sigma^2/(n/4) + \sigma^2/(n/4)} = 4\sigma/\sqrt{n}$. By using the same σ here as in the earlier calculation, we are assuming that the residual standard deviation is unchanged (or essentially unchanged) after including the interaction in the model; that is, we are assuming that inclusion of the interaction does not change R^2 much.

To put it another way, to be able to estimate the interaction to the same level of accuracy as the main effect, we would need four times the sample size.

What is the power of the estimate of the interaction, as estimated from the original experiment of size n? The probability of seeing a difference that is "statistically significant" at the 5% level is the probability that the z-score exceeds 1.96; that is, `pnorm(1.4,1.96,1) = 0.29`. And, if you do perform the analysis and report it if the 95% interval excludes zero, you will overestimate the size of the interaction by a lot, as we can see by simulating a million runs of the experiment:

```
raw <- rnorm(1e6, 1.4, 1)
significant <- raw > 1.96
mean(raw[significant])
```

The result is 2.6, implying that, on average, a statistically significant result will overestimate the size of the interaction by a factor of 2.6.

This implies a big problem with the common plan of designing a study with a focus on the main effect and then looking to see what shows up in the interactions. Or, even worse, designing a study, not finding the anticipated main effect, and then using the interactions to bail you out. The problem is

not just that this sort of analysis is "exploratory"; it's that these data are a lot noisier than you realize, so what you think of as interesting exploratory findings could be just a bunch of noise.

You need 16 times the sample size to estimate an interaction that is half the size as the main effect

As demonstrated above, if an interaction is the same size as the main effect—for example, a treatment effect of 0.5 among women, 1.5 among men, and 1.0 overall—then it will require four times the sample size to estimate with the same accuracy from a balanced experiment.

There are cases where main effects are small and interactions are large. Indeed, in general, these labels have some arbitrariness to them; for example, when studying U.S. congressional elections, recode the outcome from Democratic or Republican vote share to incumbent party vote share, and interactions with incumbent party become main effects, and main effects become interactions. So the above analysis is in the context of main effects that are modified by interactions; there's the implicit assumption that if the main effect is positive, then it will be positive in the subgroups we look at, just maybe a bit larger or smaller.

It makes sense, where possible, to code variables in a regression so that the larger comparisons appear as main effects and the smaller comparisons appear as interactions. The very nature of a "main effect" is that it is supposed to tell as much of the story as possible. When interactions are important, they are important as modifications of some main effect. This is not always the case—for example, you could have a treatment that flat-out hurts men while helping women—but in such examples it's not clear that the main-effects-plus-interaction framework is the best way of looking at things.

When a large number of interactions are being considered, we would expect most interactions to be smaller than the main effect. Consider a treatment that could interact with many possible individual characteristics, including age, sex, education, health status, and so forth. We would not expect all or most of the interactions of treatment effect with these variables to be large. Thus, when considering the challenge of estimating interactions that are not chosen ahead of time, it could be more realistic to suppose something like half the size of main effects. In that case—for example, a treatment effect of 0.75 in one group and 1.25 in the other—one would need 16 times the sample size to estimate the interaction with the same relative precision as is needed to estimate the main effect.

The message we take from this analysis is *not* that interactions are too difficult to estimate and should be ignored. Rather, interactions can be important; we just need to accept that in many settings we won't be able to attain anything like near-certainty regarding the magnitude or even direction of particular interactions. It is typically not appropriate to aim for "statistical significance" or 95% intervals that exclude zero, and it often will be appropriate to use prior information to get more stable and reasonable estimates, and to accept uncertainty, not acting as if interactions of interest are zero just because their estimate is not statistically significant.

Understanding the problem by simulating regressions in R

We can play around in R to get a sense of how standard errors for main effects and interactions depend on parameterization. For simplicity, all our simulations assume that the true (underlying) coefficients are 0. In this case, the true values are irrelevant for our goal of computing the standard error.

Example: Simulation of main effects and interactions

We start with a basic model in which we simulate 1000 data points with two predictors, each taking on the value -0.5 or 0.5. This is the same as the model above: the estimated main effects are simple differences, and the estimated interaction is a difference in differences. We also have assumed the two predictors are independent, which is what would happen in a randomized experiment where, on average, the treatment and control groups would each be expected to be evenly divided between men and women. Here is the simulation:[1]

[1]Code for this example is in the folder `SampleSize`.

```
n <- 1000
sigma <- 10
y <- rnorm(n, 0, sigma)
x1 <- sample(c(-0.5,0.5), n, replace=TRUE)
x2 <- sample(c(-0.5,0.5), n, replace=TRUE)
fake <- data.frame(c(y,x1,x2))
fit_1 <- stan_glm(y ~ x1, data=fake)
fit_2 <- stan_glm(y ~ x1 + x2 + x1:x2, data=fake)
print(fit_1)
print(fit_2)
```

And here is the result:

```
            Median MAD_SD
(Intercept) -0.1    0.3
x1           0.7    0.6
x2           0.8    0.6
x1:x2        1.2    1.3

Auxiliary parameter(s):
      Median MAD_SD
sigma 10.0    0.2
```

Ignore the estimates; they're pure noise. Just look at the standard errors. They go just as in the above formulas: $2\sigma/\sqrt{n} = 2 * 10/\sqrt{1000} = 0.6$, and $4\sigma/\sqrt{n} = 1.3$.

Now let's do the exact same thing but make the predictors take on the values 0 and 1 rather than -0.5 and 0.5:

```
fake$x1 <- sample(c(0,1), n, replace=TRUE)
fake$x2 <- sample(c(0,1), n, replace=TRUE)
fit_1 <- stan_glm(y ~ x1, data=fake)
fit_2 <- stan_glm(y ~ x1 + x2 + x1:x2, data=fake)
print(fit_1)
print(fit_2)
```

And this is what happens:

```
            Median MAD_SD
(Intercept) -0.1    0.6
x1           1.0    0.9
x2           0.1    0.9
x1:x2       -1.9    1.3

Auxiliary parameter(s):
      Median MAD_SD
sigma 10.0    0.2
```

Again, just look at the standard errors. The standard error for the interaction is still 1.3, but the standard errors for the main effects went up to 0.9. What happened?

What happened was that the main effects are now estimated at the edge of the data: the estimated coefficient of x_1 is now the difference in y, comparing the two values of x_1, just at $x_2 = 0$. So its standard error is $\sqrt{\sigma^2/(n/4) + \sigma^2/(n/4)} = 2\sqrt{2}\sigma/\sqrt{n}$. Under this parameterization, the coefficient of x_1 is estimated just from the half of the data for which $x_2 = 0$, so the standard error is $\sqrt{2}$ times as big as before. Similarly for x_2.

But these aren't really "main effects"; in the context of the above problem, the main effect of the treatment is the average over men and women. If we put the problem in a regression framework, we should be coding the predictors not as $0, 1$ but as $-0.5, 0.5$, so that the main effect for each predictor corresponds to the other predictor set to its average level.

But here's another possibility: What about coding each predictor as $-1, 1$? Let's take a look:

```
fake$x1 <- sample(c(-1,1), n, replace=TRUE)
fake$x2 <- sample(c(-1,1), n, replace=TRUE)
fit_1 <- stan_glm(y ~ x1, data=fake)
fit_2 <- stan_glm(y ~ x1 + x2 + x1:x2, data=fake)
print(fit_1)
print(fit_2)
```

This yields:

```
            Median MAD_SD
(Intercept) -0.4    0.3
x1          -0.5    0.3
x2           0.0    0.3
x1:x2        0.7    0.3

Auxiliary parameter(s):
      Median MAD_SD
sigma 9.9    0.2
```

Again, ignore the coefficient estimates and look at the standard errors. Compared to the fitted model with the $-0.5, 0.5$ coding on page 303, the standard errors for the main effects are smaller by a factor of 2, and now the standard error for the interaction has been divided by 4. What happened in this simulation?

The factor of 2 for the main effect is clear enough: If you multiply x by 2, and $\beta * x$ doesn't change, then you have to divide β by 2 to compensate, and its standard error gets divided by 2 as well. But what happened to the interaction? That's clear too: we've multiplied x_1 and x_2 each by 2, so $x_1 x_2$ is multiplied by 4.

So to make sense of all these standard errors, you have to have a feel for the appropriate scale for the coefficients.

16.5 Design calculations after the data have been collected

We return to the beauty and sex ratio example, introduced in Sections 9.4 and 9.5 to demonstrate Bayesian inference. Here we attack the problem in a slightly different way using design analysis. Either way, the message is that we can use available prior information to interpret results from particular data.

Example: Beauty and sex ratio

As a result of the intrinsic interest of the topic and the availability of data from birth records, there have been many studies of factors affecting the probability of male and female births. Most have found little or no evidence of any effects, but the study described in Section 9.4 appeared to be an exception, reporting data from a survey in which attractive parents were more likely to have daughters, a finding that was then given an explanation in terms of evolutionary biology, on the grounds that physical attractiveness enhances the reproductive success of women more than that of men.

For our discussion here we shall work with the simple analysis from Section 9.4, comparing the "very attractive" parents in the survey (56% of their children were girls) to the other parents (only 44% of their children were girls). The difference was 8% with a standard error of 3%. The classical 95% interval is $[8\% \pm 2 * 3\%] = [2\%, 14\%]$, which tells us that effects as low as 2 percentage points or as high as 14 percentage are roughly consistent with the data.

The challenge is to interpret this finding in light of our knowledge from the scientific literature that any difference in sex ratios between two such groups in the population is probably much less than 0.5% (for example, the probability of a girl birth shifting from 48.5% to 49.0%).

How, then, do we account for the fact that the 95% interval for the difference is [2%, 14%], which excludes the range of plausible differences in the population? One answer is that unusual things happen: 5% events do occur 5% of the time. A longer answer is that researchers typically have many choices or "degrees of freedom" in their analysis. For example, in this particular example, survey respondents were placed in five attractiveness categories, and the published comparison was category

5 compared to categories 1–4, pooled; see Figure 9.5. But the researcher could just as well have compared categories 4–5 to categories 1–3, or compared 3–5 to 1–2, or compared 4–5 to 1–2, and so forth. Looked at this way, it's no surprise that a determined data analyst was able to find a comparison somewhere in the data for which the 95% interval was far from zero.

What, then, can be learned from the published estimate of 8%? For the present example, the standard error of 3% means that statistical significance would only happen with an estimate of at least 6% in either direction: more than 12 times larger than any true effect that could reasonably be expected based on previous research. Thus, even if the inference of an association between parental beauty and child's sex is valid for the general population, the magnitude of the estimate from a study of this size is likely to be much larger than the true effect. This is an example of a type M (magnitude) error, as defined in Section 4.4. We can also consider the possibility of type S (sign) errors, in which a statistically significant estimate is in the opposite direction of the true effect.

We may get a sense of the probabilities of these errors by considering three scenarios of studies with standard errors of 3 percentage points:

1. *True difference of zero.* If there is no correlation between parental beauty and sex ratio of children, then a statistically significant estimate will occur 5% of the time, and it will always be misleading—a type 1 error.
2. *True difference of 0.2%.* If the probability of girl births is actually 0.2 percentage points higher among beautiful than among other parents, then what might happen with an estimate whose standard error is 3%? We can do the calculation in R: the probability of the estimate being at least 6% (two standard errors away from zero, thus "statistically significant") is `1 - pnorm(6,0.2,3)`, or 0.027, and the probability of it being at least 6% in the *negative* direction is `pnorm(-6,0.2,3)`, which comes to 0.019. The type S error rate is 0.019/(0.019 + 0.027) = 42%.
   Thus, before the data were collected, we could say that if the true population difference were 0.2%, that this study has a 3% probability of being statistically significant and positive—and a 2% chance of being statistically significant negative result. If the estimate is statistically significant, it must be at least 6 percentage points, thus at least 30 times higher than the true effect, and with a 40% chance of going in the wrong direction.
3. *True difference of 0.5%.* If the probability of girl births is actually 0.5 percentage point higher among beautiful than among other parents—which, based on the literature, is well beyond the high end of possible effect sizes—then there is a 0.033 chance of a statistically significant positive result, and a 0.015 chance of a statistically significant result in the wrong direction. The type S error rate is 0.015/(0.015 + 0.033) = 31%.
   So, if the true difference is 0.5%, any statistically significant estimated effect will be at least 12 times the magnitude of the true effect and with a 30% chance of having the wrong sign. Thus, again, the experiment gives little information about the sign or the magnitude of the true effect.

A sample of this size is just not useful for estimating variation on the order of half a percentage points or less, which is why most studies of the human sex ratio use much larger samples, typically from demographic databases. The example shows that if the sample is too small relative to the expected size of any differences, it is not possible to draw strong conclusions even when estimates are seemingly statistically significant.

Indeed, with this level of noise, *only* very large estimated effects could make it through the statistical significance filter. The result is almost a machine for producing exaggerated claims, which become only more exaggerated when they hit the news media with the seal of scientific approval.

It is well known that with a large enough sample size, even a very small estimate can be statistically significantly different from zero. Many textbooks contain warnings about mistaking statistical significance in a large sample for practical importance. It is also well known that it is difficult to obtain statistically significant results in a small sample. Consequently, when results are significant despite the handicap of a small sample, it is natural to think that they are real and important. The above example shows then this is not necessarily the case.

If the estimated effects in the sample are much larger than those that might reasonably be expected

in the population, even seemingly statistically significant results provide only weak evidence of any effect. Yet one cannot simply ask researchers to avoid using small samples. There are cases in which it is difficult or impossible to obtain more data, and researchers must make do with what is available.

In such settings, researchers should determine plausible effect sizes based on previous research or theory, and carry out design calculations based on the observed test statistics. Conventional significance levels tell us how often the observed test statistic would be obtained if there were no effect, but one should also ask how often the observed test statistic would be obtained under a reasonable assumption about the size of effects. Estimates that are much larger than expected might reflect population effects that are much larger than previously imagined. Often, however, large estimates will merely reflect the influence of random variation. It may be disappointing to researchers to learn that even estimates that are both "statistically" and "practically" significant do not necessarily provide strong evidence. Accurately identifying findings that are suggestive rather than definitive, however, should benefit both the scientific community and the general public.

16.6 Design analysis using fake-data simulation

Example: Fake-data simulation for experimental design

The most general and often the clearest method for studying the statistical properties of a proposed design is to simulate the data that might be collected along with the analyses that could be performed. We demonstrate with an artificial example of a randomized experiment on 100 students designed to test an intervention for improving final exam scores.[2]

Simulating a randomized experiment

We start by assigning the potential outcomes, the final exam scores that would be observed for each student if he or she gets the control or the treatment:

```
n <- 100
y_if_control <- rnorm(n, 60, 20)
y_if_treated <- y_if_control + 5
```

In this very simple model, the intervention would add 5 points to each student's score.

We then assign treatments ($z = 0$ for control or 1 for treatment), which then determine which outcome is observed for each person:

```
z <- sample(rep(c(0,1), n/2))
y <- ifelse(z==1, y_if_treated, y_if_control)
fake <- data.frame(y, z)
```

Having simulated the data, we can now compare treated to control outcomes and compute the standard error for the difference:

```
diff <- mean(y[z==1]) - mean(y[z==0])
se_diff <- sqrt(sd(y[z==0])^2/sum(z==0) + sd(y[z==1])^2/sum(z==1))
```

Equivalently (see Section 7.3), we can run the regression:

```
fit_1a <- stan_glm(y ~ z, data=fake)
```

which yields,

```
            Median MAD_SD
(Intercept) 66.0    2.7
z           -2.8    4.3

Auxiliary parameter(s):
      Median MAD_SD
sigma 21.3    1.5
```

[2]Code for this example is in the folder FakeMidtermFinal.

The parameter of interest here is the coefficient of z, and its standard error is 4.3, suggesting that, under these conditions, a sample size of 100 would not be enough to get a good estimate of a treatment effect of 5 points. The standard error of 4.3 is fairly precisely estimated, as we can tell because the uncertainty in `sigma` is low compared to its estimate.

When looking at the above simulation result to assess this design choice, we should focus on the standard error of the parameter of interest (in this case, 4.0) and compare it to the assumed parameter value (in this case, 5), *not* to the noisy point estimate from the simulation (in this case -2.8).

To give a sense of why it would be a mistake to focus on the point estimate, we repeat the above steps, simulating for a new batch of 100 students simulated from the model. Here is the result:

```
(Intercept) 59.7    2.9
z           11.8    4.0

Auxiliary parameter(s):
      Median MAD_SD
sigma 20.1    1.4
```

A naive read of this table would be that the design with 100 students is just fine, as the estimate is well over two standard errors away from zero. But that conclusion would be a mistake, as the coefficient estimate here is too noisy to be useful.

The above simulation indicates that, under the given assumptions, the randomized design with 100 students gives an estimate of the treatment effect with standard error of approximately 4 points. If that is acceptable, fine. If not, one approach would be to increase the sample size. Standard error decreases with the square root of sample size, so if, for example, we wanted to reduce the standard error to 2 points, we would need a sample size of approximately 400.

Including a pre-treatment predictor

Another approach to increase efficiency is to consider a pre-test. Suppose pre-test scores x have the same distribution as post-test scores y but with a slightly lower average:

```
fake$x <- rnorm(n, 50, 20)
```

We can then adjust for pre-test in our regression:

```
fit_1b <- stan_glm(y ~ z + x, data=fake)

            Median MAD_SD
(Intercept) 51.3    5.9
z           10.9    4.5
x            0.2    0.1

Auxiliary parameter(s):
      Median MAD_SD
sigma 21.1    1.5
```

Again, the coefficient of z estimates the treatment effect, and it still has a standard error of about 4, which might seem surprising: shouldn't the inclusion of a pre-treatment predictor increase the precision of our estimate? The answer is that, the way we constructed the pre-test variable, it wasn't much of a pre-treatment predictor at all, as we simulated it independently of the potential outcomes for the final test score.

To perform a realistic simulation, we must simulate both test scores in a correlated way, which we do here by borrowing a trick from the example of simulated midterm and final exams in Section 6.5:

1. Each student is assumed to have a true ability drawn from a distribution with mean 50 and standard deviation 16.

2. Each student's score on the pre-test, x, is the sum of two components: the student's true ability, and a random component with mean 0 and standard deviation 12, reflecting that performance on any given test will be unpredictable.
3. Each student's score on the post-test, y, is his or her true ability, plus another, independent, random component, plus an additional 10 points if a student receives the control or 15 points if he or she receives the treatment.

These are the same conditions as in Section 6.5, except that (i) we have increased the standard deviations of each component of the model so that the standard deviation of the final scores, $\sqrt{16^2 + 12^2} = 20$, is consistent with the distribution assumed for y in our simulations above, and (ii) we have increased the average score level on the post-test along with a treatment effect.

Here is the code to create the artificial world:

```
n <- 100
true_ability <- rnorm(n, 50, 16)
x <- true_ability + rnorm(n, 0, 12)
y_if_control <- true_ability + rnorm(n, 0, 12) + 10
y_if_treated <- y_if_control + 5
```

As above, we assign treatments, construct the observed outcome, and put the data into a frame:

```
z <- sample(rep(c(0,1), n/2))
y <- ifelse(z==1, y_if_treated, y_if_control)
fake_2 <- data.frame(x, y, z)
```

The simple comparison is equivalent to a regression on the treatment indicator:

```
fit_2a <- stan_glm(y ~ z, data=fake_2)

            Median MAD_SD
(Intercept) 59.2    3.0
z            9.5    4.3

Auxiliary parameter(s):
      Median MAD_SD
sigma 21.6    1.5
```

And the estimate adjusting for pre-test:

```
fit_2b <- stan_glm(y ~ z + x, data=fake_2)

            Median MAD_SD
(Intercept) 27.4    4.3
z            6.1    3.3
x            0.7    0.1

Auxiliary parameter(s):
      Median MAD_SD
sigma 16.2    1.2
```

In this case, with the strong dependence between pre-test and post-test, this adjustment has reduced the residual standard deviation by about a third.

Simulating an experiment with selection bias

With data coming from a randomized experiment, all the regressions considered above give unbiased estimates of the treatment effect. But suppose we are concerned about bias in the treatment assignment. We can simulate that too.

For example, suppose that school administrators, out of kindness, are more likely to give the

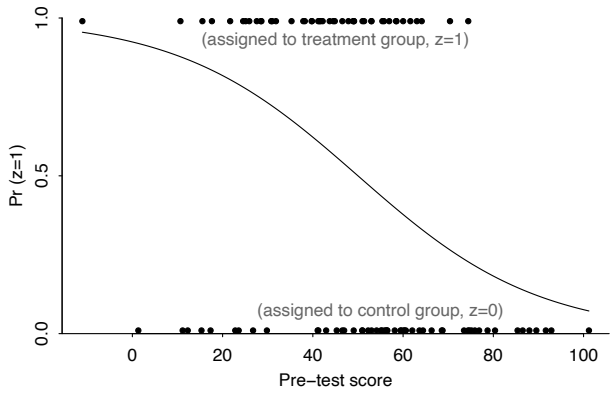

Figure 16.5 *Simulated treatment assignments based on a rule in which students with lower pre-test scores are more likely to get the treatment. We use this to demonstrate how a simulation study can be used to assess the bias in a design and estimation procedure.*

treatment to students who are performing poorly. We could simulate this behavior with an unequal-probability assignment rule such as $\Pr(z_i = 1) = \text{logit}^{-1}(-(x_i - 50)/20)$, where we have chosen the logistic curve for convenience and set its parameters so that the probability averages to approximately 0.5, with a bit of variation from one end of the data to the other. Figure 16.5 shows the assumed logistic curve and the simulated treatment assignments for the 100 students in this example, as produced by the following code:

```
z <- rbinom(n, 1, invlogit(-(x-50)/20))
```

We then record the observed post-test and save as a data frame:

```
y <- ifelse(z==1, y_if_treated, y_if_control)
fake_3 <- data.frame(x, y, z)
```

By construction, the true treatment effect is 5 points, as before, but a simple comparison yields a biased estimate, while the linear regression adjusting for pre-test is better.

To see this, we should not just perform one simulation; as discussed earlier in this section, not much can be learned from the estimate obtained from any single simulation. Instead we first write a function to simulate the fake data, assign the treatments, and perform the simple comparison and the regression adjusting for pre-test:

```
experiment <- function(n) {
  true_ability <- rnorm(n, 50, 16)
  x <- true_ability + rnorm(n, 0, 12)
  y_if_control <- true_ability + rnorm(n, 0, 12) + 10
  y_if_treated <- y_if_control + 5
  z <- rbinom(n, 1, invlogit(-(x-50)/20))
  y <- ifelse(z==1, y_if_treated, y_if_control)
  fake_3 <- data.frame(x, y, z)
  fit_3a <- stan_glm(y ~ z, data=fake_3, refresh=0)
  fit_3b <- stan_glm(y ~ z + x, data=fake_3, refresh=0)
  rbind(c(coef(fit_3a)["z"], se(fit_3a)["z"]), c(coef(fit_3b)["z"], se(fit_3b)["z"]))
}
```

We then loop this simulation 50 times:

```
n <- 100
n_loop <- 50
results <- array(NA, c(n_loop, 2, 2),
  dimnames=list(1:n_loop, c("Simple", "Adjusted"), c("Estimate", "SE")))
  for (loop in 1:n_loop){
```

```
  results[loop,,] <- experiment(n)
}
```

The above steps produce a $50 \times 2 \times 2$ matrix which we then average over to compute a 2×2 matrix of average estimate and average standard error for the two procedures:

```
results_avg <- apply(results, c(2,3), mean)
```

Here is the result:

```
          Estimate  SE
Simple       -6.4 3.9
Adjusted      4.6 3.4
```

The true parameter value here is 5.0, so in this case the simple comparison is horribly biased—no surprise if you reflect upon the big differences between treatment and control groups from the simulation shown in Figure 16.5. In contrast, the bias of the adjusted estimate is low. In other settings, for example if the underlying relation between pre-test and post-test is nonlinear, or if there is selection on an unobserved or unmodeled variable, the regression-adjusted estimate can have a large bias too. We discuss these topics further in Chapters 18–21; our point here is that you can assess such biases using simulation, conditional on a model for data, measurement, and treatment assignment.

16.7 Bibliographic note

The quote at the beginning of Section 16.1 is from Weisburd, Petrosino, and Mason (1993); see also Gelman, Skardhamar, and Aaltonen (2017). The problems of statistical power are discussed by Button et al. (2013) and Gelman (2019a). Figure 16.1 comes from Gelman (2015d).

Cochran (1977) and Lohr (2009) are standard and useful references for classical models in survey sampling. Groves et al. (2009) and Heeringa, West, and Berglund (2017) go over practical aspects of survey design and analysis. Yates (1967), Montgomery (1986), and Box, Hunter, and Hunter (2005) review the statistical aspects of experimental design.

Hoenig and Heisey (2001), Lenth (2001), and Gelman and Carlin (2014) provide some general warnings and advice on sample size and power calculations. Assmann et al. (2000) discuss the general difficulty of estimating interactions. The design calculations for the sex ratio example in Section 16.5 are taken from Gelman and Weakliem (2009).

16.8 Exercises

16.1 *Sample size calculations for estimating a proportion*:

(a) How large a sample survey would be required to estimate, to within a standard error of $\pm$ 3%, the proportion of the U.S. population who support the death penalty?

(b) About 14% of the U.S. population is Latino. How large would a national sample of Americans have to be in order to estimate, to within a standard error of $\pm$3%, the proportion of Latinos in the United States who support the death penalty?

(c) How large would a national sample of Americans have to be in order to estimate, to within a standard error of $\pm$ 1%, the proportion who are Latino?

16.2 *Sample size calculation for estimating a difference*: Consider an election with two major candidates, A and B, and a minor candidate, C, who are believed to have support of approximately 45%, 35%, and 20% in the population. A poll is to be conducted with the goal of estimating the difference in support between candidates A and B. How large a sample would you estimate is needed to estimate this difference to within a standard error of 5 percentage points? (Hint: consider an outcome variable that is coded as $+1$, -1, and 0 for supporters of A, B, and C, respectively.)

16.3 *Power*: Following Figure 16.3, determine the power (the probability of getting an estimate that is "statistically significantly" different from zero at the 5% level) of a study where the true effect size is X standard errors from zero. Answer for the following values of X: 0, 1, 2, and 3.

16.4 *Power, type M error, and type S error*: Consider the experiment shown in Figure 16.1 where the true effect could not realistically be more than 2 percentage points and it is estimated with a standard error of 8.1 percentage points.

(a) Assuming the estimate is unbiased and normally distributed and the true effect size is 2 percentage points, use simulation to answer the following questions: What is the power of this study? What is the type M error rate? What is the type S error rate?

(b) Assuming the estimate is unbiased and normally distributed and the true effect size is *no more than* 2 percentage points in absolute value, what can you say about the power, type M error rate, and type S error rate?

16.5 *Design analysis for an experiment*: You conduct an experiment in which half the people get a special get-out-the-vote message and others do not. Then you follow up after the election with a random sample of 500 people to see if they voted.

(a) What will be the standard error of your estimate of effect size? Figure this out making reasonable assumptions about voter turnout and the true effect size.

(b) Check how sensitive your standard error calculation is to your assumptions.

(c) For a range of plausible effect sizes, consider conclusions from this study, in light of the statistical significance filter. As a researcher, how can you avoid this problem?

16.6 *Design analysis with pre-treatment information*: A new teaching method is hoped to increase scores by 5 points on a certain standardized test. An experiment is performed on n students, where half get this intervention and half get the control. Suppose that the standard deviation of test scores in the population is 20 points. Further suppose that a pre-test is available which has a correlation of 0.8 with the post-test under the control condition. What will be the standard error of the estimated treatment effect based on a fitted regression, assuming that the treatment effect is constant and independent of the value of the pre-test?

16.7 *Decline effect*: After a study is published on the effect of some treatment or intervention, it is common for the estimated effect in future studies to be lower. Give five reasons why you might expect this to happen.

16.8 *Effect size and sample size*: Consider a toxin that can be tested on animals at different doses. Suppose a typical exposure level for humans is 1 (in some units), and at this level the toxin is hypothesized to introduce a risk of 0.01% of death per person.

(a) Consider different animal studies, each time assuming a linear dose-response relation (that is, 0.01% risk of death per animal per unit of the toxin), with doses of 1, 100, and 10 000. At each of these exposure levels, what sample size is needed to have 80% power of detecting the effect?

(b) This time assume that response is a logged function of dose and redo the calculations in (a).

16.9 *Cluster sampling with equal-sized clusters*: A survey is being planned with the goal of interviewing n people in some number J of clusters. For simplicity, assume simple random sampling of clusters and a simple random sample of size n/J (appropriately rounded) within each sampled cluster.

Consider inferences for the proportion of Yes responses in the population for some question of interest. The estimate will be simply the average response for the n people in the sample. Suppose that the true proportion of Yes responses is not too far from 0.5 and that the standard deviation among the mean responses of clusters is 0.1.

(a) Suppose the total sample size is $n = 1000$. What is the standard error for the sample average if $J = 1000$? What if $J = 100$, 10, 1?

(b) Suppose the cost of the survey is $50 per interview, plus $500 per cluster. Further suppose that the goal is to estimate the proportion of Yes responses in the population with a standard error of no more than 2%. What values of n and J will achieve this at the lowest cost?

16.10 *Simulation for design analysis*: The folder `ElectricCompany` contains data from the Electric Company experiment analyzed in Chapter 19. Suppose you wanted to perform a new experiment under similar conditions, but for simplicity just for second graders, with the goal of having 80% power to find a statistically significant result (at the 5% level) in grade 2.

(a) State clearly the assumptions you are making for your design calculations. (Hint: you can set the numerical values for these assumptions based on the analysis of the existing Electric Company data.)

(b) Suppose that the new data will be analyzed by simply comparing the average scores for the treated classrooms to the average scores for the controls. How many classrooms would be needed for 80% power?

(c) Repeat (b), but supposing that the new data will be analyzed by comparing the average gain scores for the treated classrooms to the average gain scores of the controls.

(d) Repeat, but supposing that the new data will be analyzed by regression, adjusting for pre-test scores as well as the treatment indicator.

16.11 *Optimal design*:

(a) Suppose that the zinc study described in Section 16.3 would cost $150 for each treated child and $100 for each control. Under the assumptions given in that section, determine the number of control and treated children needed to attain 80% power at minimal total cost. You will need to set up a loop of simulations as illustrated for the example in the text. Assume that the number of measurements per child is fixed at $K = 7$ (that is, measuring every two months for a year).

(b) Make a generalization of Figure 16.1 with several lines corresponding to different values of the design parameter K, the number of measurements for each child.

16.12 *Experiment with pre-treatment information:* An intervention is hoped to increase voter turnout in a local election from 20% to 25%.

(a) In a simple randomized experiment, how large a sample size would be needed so that the standard error of the estimated treatment effect is less than 2 percentage points?

(b) Now suppose that previous voter turnout was known for all participants in the experiment. Make a reasonable assumption about the correlation between turnout in two successive elections. Under this assumption, how much would the standard error decrease if previous voter turnout was included as a pre-treatment predictor in a regression to estimate the treatment effect?

16.13 *Sample size calculations for main effects and interactions*: In causal inference, it is often important to study varying treatment effects: for example, a treatment could be more effective for men than for women, or for healthy than for unhealthy patients. Suppose a study is designed to have 80% power to detect a main effect at a 95% confidence level. Further suppose that interactions of interest are half the size of main effects.

(a) What is its power for detecting an interaction, comparing men to women (say) in a study that is half men and half women?

(b) Suppose 1000 studies of this size are performed. How many of the studies would you expect to report a "statistically significant" interaction? Of these, what is the expectation of the ratio of estimated effect size to actual effect size?

16.14 *Working through your own example*: Continuing the example from the final exercises of the earlier chapters, think of a new data survey, experiment, or observational study that could be relevant and perform a design analysis for it, addressing issues of measurement, precision, and sample size. Simulate fake data for this study and analyze the simulated data.