**f.** Using the test of part (e), what would you conclude from the following sample data:

9.981 10.006 9.857 10.107 9.888
9.728 10.439 10.214 10.190 9.793

**g.** Reexpress the test procedure of part (b) in terms of the standardized test statistic $Z = (\bar{X} - 10)/(\sigma/\sqrt{n})$.

**12.** A new design for the braking system on a certain type of car has been proposed. For the current system, the true average braking distance at 40 mph under specified conditions is known to be 120 ft. It is proposed that the new design be implemented only if sample data strongly indicates a reduction in true average braking distance for the new design.

**a.** Define the parameter of interest and state the relevant hypotheses.

**b.** Suppose braking distance for the new system is normally distributed with $\sigma = 10$. Let $\bar{X}$ denote the sample average braking distance for a random sample of 36 observations. Which of the following rejection regions is appropriate: $R_1 = \{\bar{x}: \bar{x} \geq 124.80\}$, $R_2 = \{\bar{x}: \bar{x} \leq 115.20\}$, $R_3 = \{\bar{x}:$ either $\bar{x} \geq 125.13$ or $\bar{x} \leq 114.87\}$?

**c.** What is the significance level for the appropriate region of part (b)? How would you change the region to obtain a test with $\alpha = .001$?

**d.** What is the probability that the new design is not implemented when its true average braking distance is actually 115 ft and the appropriate region from part (b) is used?

**e.** Let $Z = (\bar{X} - 120)/(\sigma/\sqrt{n})$. What is the significance level for the rejection region $\{z: z \leq -2.33\}$? For the region $\{z: z \leq -2.88\}$?

**13.** Let $X_1, \ldots, X_n$ denote a random sample from a normal population distribution with a known value of $\sigma$.

**a.** For testing the hypotheses $H_0$: $\mu = \mu_0$ versus $H_a$: $\mu > \mu_0$ (where $\mu_0$ is a fixed number), show that the test with test statistic $\bar{X}$ and rejection region $\bar{x} \geq \mu_0 + 2.33\sigma/\sqrt{n}$ has significance level .01.

**b.** Suppose the procedure of part (a) is used to test $H_0$: $\mu \leq \mu_0$ versus $H_a$: $\mu > \mu_0$. If $\mu_0 = 100$, $n = 25$, and $\sigma = 5$, what is the probability of committing a type I error when $\mu = 99$? When $\mu = 98$? In general, what can be said about the probability of a type I error when the actual value of $\mu$ is less than $\mu_0$? Verify your assertion.

**14.** Reconsider the situation of Exercise 11 and suppose the rejection region is $\{\bar{x}: \bar{x} \geq 10.1004$ or $\bar{x} \leq 9.8940\} = \{z: z \geq 2.51$ or $z \leq -2.65\}$.

**a.** What is $\alpha$ for this procedure?

**b.** What is $\beta$ when $\mu = 10.1$? When $\mu = 9.9$? Is this desirable?

## 8.2 Tests About a Population Mean

The general discussion in Chapter 7 of confidence intervals for a population mean $\mu$ focused on three different cases. We now develop test procedures for these same three cases.

### Case I: A Normal Population with Known $\sigma$

Although the assumption that the value of $\sigma$ is known is rarely met in practice, this case provides a good starting point because of the ease with which general procedures and their properties can be developed. The null hypothesis in all three cases will state that $\mu$ has a particular numerical value, the *null value,* which we will denote by $\mu_0$. Let $X_1, \ldots, X_n$ represent a random sample of size $n$ from the normal population. Then the sample mean $\bar{X}$ has a normal distribution with expected value $\mu_{\bar{X}} = \mu$ and standard deviation $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. When $H_0$ is true, $\mu_{\bar{X}} = \mu_0$. Consider now the statistic $Z$ obtained by standardizing $\bar{X}$ under the assumption that $H_0$ is true:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

Substitution of the computed sample mean $\bar{x}$ gives $z$, the distance between $\bar{x}$ and $\mu_0$ expressed in "standard deviation units." For example, if the null hypothesis is $H_0$: $\mu = 100$, $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 10/\sqrt{25} = 2.0$ and $\bar{x} = 103$, then the test statistic value is $z = (103 - 100)/2.0 = 1.5$. That is, the observed value of $\bar{x}$ is 1.5 standard deviations (of $\bar{X}$) above what we expect it to be when $H_0$ is true. The statistic $Z$ is a natural measure of the distance between $\bar{X}$, the estimator of $\mu$, and its expected value when $H_0$ is true. If this distance is too great in a direction consistent with $H_a$, the null hypothesis should be rejected.

Suppose first that the alternative hypothesis has the form $H_a$: $\mu > \mu_0$. Then an $\bar{x}$ value less than $\mu_0$ certainly does not provide support for $H_a$. Such an $\bar{x}$ corresponds to a negative value of $z$ (since $\bar{x} - \mu_0$ is negative and the divisor $\sigma/\sqrt{n}$ is positive). Similarly, an $\bar{x}$ value that exceeds $\mu_0$ by only a small amount (corresponding to $z$ which is positive but small) does not suggest that $H_0$ should be rejected in favor of $H_a$. The rejection of $H_0$ is appropriate only when $\bar{x}$ considerably exceeds $\mu_0$—that is, when the $z$ value is positive and large. In summary, the appropriate rejection region, based on the test statistic $Z$ rather than $\bar{X}$, has the form $z \geq c$.

As discussed in Section 8.1, the cutoff value $c$ should be chosen to control the probability of a type I error at the desired level $\alpha$. This is easily accomplished because the distribution of the test statistic $Z$ when $H_0$ is true is the standard normal distribution (that's why $\mu_0$ was subtracted in standardizing). The required cutoff $c$ is the $z$ critical value that captures upper-tail area $\alpha$ under the standard normal curve. As an example, let $c = 1.645$, the value that captures tail area .05 ($z_{.05} = 1.645$). Then,

$$\begin{aligned}\alpha &= P(\text{type I error}) = P(H_0 \text{ is rejected when } H_0 \text{ is true}) \\ &= P(Z \geq 1.645 \text{ when } Z \sim N(0, 1)) = 1 - \Phi(1.645) = .05\end{aligned}$$

More generally, the rejection region $z \geq z_\alpha$ has type I error probability $\alpha$. The test procedure is *upper-tailed* because the rejection region consists only of large values of the test statistic.

Analogous reasoning for the alternative hypothesis $H_a$: $\mu < \mu_0$ suggests a rejection region of the form $z \leq c$, where $c$ is a suitably chosen negative number ($\bar{x}$ is far below $\mu_0$ if and only if $z$ is quite negative). Because $Z$ has a standard normal distribution when $H_0$ is true, taking $c = -z_\alpha$ yields $P(\text{type I error}) = \alpha$. This is a *lower-tailed* test. For example, $z_{.10} = 1.28$ implies that the rejection region $z \leq -1.28$ specifies a test with significance level .10.

Finally, when the alternative hypothesis is $H_a$: $\mu \neq \mu_0$, $H_0$ should be rejected if $\bar{x}$ is too far to either side of $\mu_0$. This is equivalent to rejecting $H_0$ either if $z \geq c$ or if $z \leq -c$. Suppose we desire $\alpha = .05$. Then,

$$\begin{aligned}.05 &= P(Z \geq c \text{ or } Z \leq -c \text{ when } Z \text{ has a standard normal distribution}) \\ &= \Phi(-c) + 1 - \Phi(c) = 2[1 - \Phi(c)]\end{aligned}$$

Thus $c$ is such that $1 - \Phi(c)$, the area under the standard normal curve to the right of $c$, is .025 (and not .05!). From Section 4.3 or Appendix Table A.3, $c = 1.96$, and the rejection region is $z \geq 1.96$ *or* $z \leq -1.96$. For any $\alpha$, the *two-tailed* rejection region $z \geq z_{\alpha/2}$ *or* $z \leq -z_{\alpha/2}$ has type I error probability $\alpha$ (since area $\alpha/2$ is captured under each of the two tails of the $z$ curve). Again, the key reason for using the standardized test

statistic $Z$ is that because $Z$ has a known distribution when $H_0$ is true (standard normal), a rejection region with desired type I error probability is easily obtained by using an appropriate critical value.

The test procedure for case I is summarized in the accompanying box, and the corresponding rejection regions are illustrated in Figure 8.2.

Null hypothesis: $H_0$: $\mu = \mu_0$

Test statistic value: $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$

| Alternative Hypothesis | Rejection Region for Level $\alpha$ Test |
|---|---|
| $H_a$: $\mu > \mu_0$ | $z \geq z_\alpha$ (upper-tailed test) |
| $H_a$: $\mu < \mu_0$ | $z \leq -z_\alpha$ (lower-tailed test) |
| $H_a$: $\mu \neq \mu_0$ | either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$ (two-tailed test) |

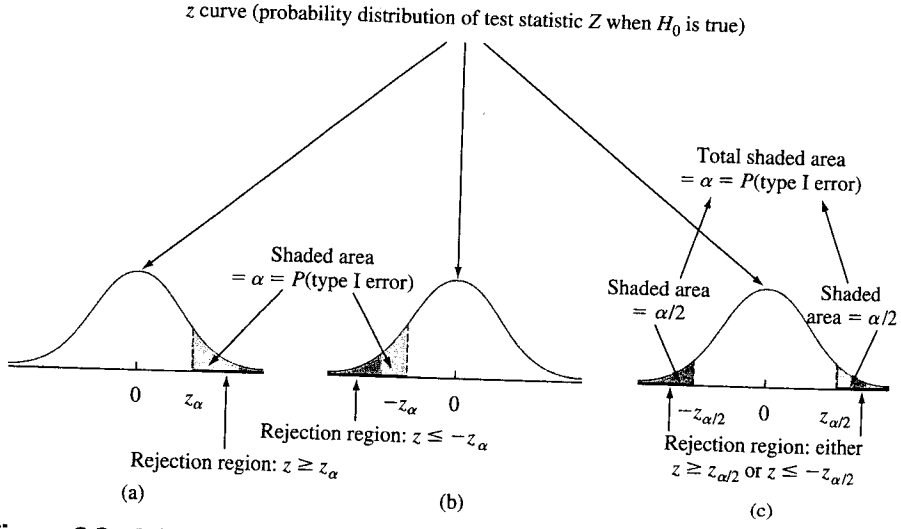


**Figure 8.2** Rejection regions for $z$ tests: (a) upper-tailed test; (b) lower-tailed test; (c) two-tailed test

Use of the following sequence of steps is recommended when testing hypotheses about a parameter.

1. Identify the parameter of interest and describe it in the context of the problem situation.
2. Determine the null value and state the null hypothesis.
3. State the appropriate alternative hypothesis.

4. Give the formula for the computed value of the test statistic (substituting the null value and the known values of any other parameters, but *not* those of any sample-based quantities).
5. State the rejection region for the selected significance level $\alpha$.
6. Compute any necessary sample quantities, substitute into the formula for the test statistic value, and compute that value.
7. Decide whether $H_0$ should be rejected and state this conclusion in the problem context.

The formulation of hypotheses (Steps 2 and 3) should be done before examining the data.

**Example 8.6** A manufacturer of sprinkler systems used for fire protection in office buildings claims that the true average system-activation temperature is 130°. A sample of $n = 9$ systems, when tested, yields a sample average activation temperature of 131.08°F. If the distribution of activation times is normal with standard deviation 1.5°F, does the data contradict the manufacturer's claim at significance level $\alpha = .01$?

1. Parameter of interest: $\mu$ = true average activation temperature.
2. Null hypothesis: $H_0$: $\mu = 130$ (null value = $\mu_0 = 130$).
3. Alternative hypothesis: $H_a$: $\mu \neq 130$ (a departure from the claimed value in *either* direction is of concern).
4. Test statistic value:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x} - 130}{1.5/\sqrt{n}}$$

5. Rejection region: The form of $H_a$ implies use of a two-tailed test with rejection region *either* $z \geq z_{.005}$ *or* $z \leq -z_{.005}$. From Section 4.3 or Appendix Table A.3, $z_{.005} = 2.58$, so we reject $H_0$ if either $z \geq 2.58$ or $z \leq -2.58$.
6. Substituting $n = 9$ and $\bar{x} = 131.08$,

$$z = \frac{131.08 - 130}{1.5/\sqrt{9}} = \frac{1.08}{.5} = 2.16$$

That is, the observed sample mean is a bit more than 2 standard deviations above what would have been expected were $H_0$ true.

7. The computed value $z = 2.16$ does not fall in the rejection region ($-2.58 < 2.16 < 2.58$), so $H_0$ cannot be rejected at significance level .01. The data does not give strong support to the claim that the true average differs from the design value of 130. ■

**$\beta$ and Sample Size Determination** The $z$ tests for case I are among the few in statistics for which there are simple formulas available for $\beta$, the probability of a type II error. Consider first the upper-tailed test with rejection region $z \geq z_\alpha$. This is equivalent to $\bar{x} \geq \mu_0 + z_\alpha \cdot \sigma/\sqrt{n}$, so $H_0$ will not be rejected if $\bar{x} < \mu_0 + z_\alpha \cdot \sigma/\sqrt{n}$. Now let $\mu'$ denote a particular value of $\mu$ that exceeds the null value $\mu_0$. Then,

$$\begin{aligned}\beta(\mu') &= P(H_0 \text{ is not rejected when } \mu = \mu') \\ &= P(\bar{X} < \mu_0 + z_\alpha \cdot \sigma/\sqrt{n} \text{ when } \mu = \mu') \\ &= P\left(\frac{\bar{X} - \mu'}{\sigma/\sqrt{n}} < z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} \text{ when } \mu = \mu'\right) \\ &= \Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)\end{aligned}$$

As $\mu'$ increases, $\mu_0 - \mu'$ becomes more negative, so $\beta(\mu')$ will be small when $\mu'$ greatly exceeds $\mu_0$ (because the value at which $\Phi$ is evaluated will then be quite negative). Error probabilities for the lower-tailed and two-tailed tests are derived in an analogous manner.

If $\sigma$ is large, the probability of a type II error can be large at an alternative value $\mu'$ that is of particular concern to an investigator. Suppose we fix $\alpha$ and also specify $\beta$ for such an alternative value. In the sprinkler example, company officials might view $\mu' = 132$ as a very substantial departure from $H_0$: $\mu = 130$ and therefore wish $\beta(132) = .10$ in addition to $\alpha = .01$. More generally, consider the two restrictions $P(\text{type I error}) = \alpha$ and $\beta(\mu') = \beta$ for specified $\alpha$, $\mu'$, and $\beta$. Then for an upper-tailed test, the sample size $n$ should be chosen to satisfy

$$\Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right) = \beta$$

This implies that

$$-z_\beta = \begin{matrix}z \text{ critical value that} \\ \text{captures lower-tail area } \beta\end{matrix} = z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}$$

It is easy to solve this equation for the desired $n$. A parallel argument yields the necessary sample size for lower- and two-tailed tests as summarized in the next box.

| **Alternative Hypothesis** | **Type II Error Probability $\beta(\mu')$ for a Level $\alpha$ Test** |
|---|---|
| $H_a$: $\mu > \mu_0$ | $\Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)$ |
| $H_a$: $\mu < \mu_0$ | $1 - \Phi\left(-z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)$ |
| $H_a$: $\mu \neq \mu_0$ | $\Phi\left(z_{\alpha/2} + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right) - \Phi\left(-z_{\alpha/2} + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)$ |

where $\Phi(z)$ = the standard normal cdf.

The sample size $n$ for which a level $\alpha$ test also has $\beta(\mu') = \beta$ at the alternative value $\mu'$ is

$$n = \begin{cases} \left[\frac{\sigma(z_\alpha + z_\beta)}{\mu_0 - \mu'}\right]^2 & \text{for a one-tailed (upper or lower) test} \\ \left[\frac{\sigma(z_{\alpha/2} + z_\beta)}{\mu_0 - \mu'}\right]^2 & \text{for a two-tailed test (an approximate solution)} \end{cases}$$

**Example 8.7** Let $\mu$ denote the true average tread life of a certain type of tire. Consider testing $H_0$: $\mu = 30{,}000$ versus $H_a$: $\mu > 30{,}000$ based on a sample of size $n = 16$ from a normal population distribution with $\sigma = 1500$. A test with $\alpha = .01$ requires $z_\alpha = z_{.01} = 2.33$. The probability of making a type II error when $\mu = 31{,}000$ is

$$\beta(31{,}000) = \Phi\left(2.33 + \frac{30{,}000 - 31{,}000}{1500/\sqrt{16}}\right) = \Phi(-.34) = .3669$$

Since $z_{.1} = 1.28$, the requirement that the level .01 test also have $\beta(31{,}000) = .1$ necessitates

$$n = \left[\frac{1500(2.33 + 1.28)}{30{,}000 - 31{,}000}\right]^2 = (-5.42)^2 = 29.32$$

The sample size must be an integer, so $n = 30$ tires should be used. ■

## Case II: Large-Sample Tests

When the sample size is large, the $z$ tests for case I are easily modified to yield valid test procedures without requiring either a normal population distribution or known $\sigma$. The key result was used in Chapter 7 to justify large-sample confidence intervals: A large $n$ implies that the sample standard deviation $s$ will be close to $\sigma$ for most samples, so that the standardized variable

$$Z = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has *approximately* a standard normal distribution. Substitution of the null value $\mu_0$ in place of $\mu$ yields the test statistic

$$Z = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

which has approximately a standard normal distribution when $H_0$ is true. The use of rejection regions given previously for case I (e.g., $z \geq z_\alpha$ when the alternative hypothesis is $H_a$: $\mu > \mu_0$) then results in test procedures for which the significance level is approximately (rather than exactly) $\alpha$. The rule of thumb $n > 40$ will again be used to characterize a large sample size.

**Example 8.8** A dynamic cone penetrometer (DCP) is used for measuring material resistance to penetration (mm/blow) as a cone is driven into pavement or subgrade. Suppose that for a particular application, it is required that the true average DCP value for a certain type of pavement be less than 30. The pavement will not be used unless there is conclusive evidence that the specification has been met. Let's state and test the appropriate hypotheses using the following data ("Probabilistic Model for the Analysis of Dynamic Cone Penetrometer Test Values in Pavement Structure Evaluation," *J. of Testing and Evaluation*, 1999: 7–14):

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 14.1 | 14.5 | 15.5 | 16.0 | 16.0 | 16.7 | 16.9 | 17.1 | 17.5 | 17.8 |
| 17.8 | 18.1 | 18.2 | 18.3 | 18.3 | 19.0 | 19.2 | 19.4 | 20.0 | 20.0 |
| 20.8 | 20.8 | 21.0 | 21.5 | 23.5 | 27.5 | 27.5 | 28.0 | 28.3 | 30.0 |
| 30.0 | 31.6 | 31.7 | 31.7 | 32.5 | 33.5 | 33.9 | 35.0 | 35.0 | 35.0 |
| 36.7 | 40.0 | 40.0 | 41.3 | 41.7 | 47.5 | 50.0 | 51.0 | 51.8 | 54.4 |
| 55.0 | 57.0 | | | | | | | | |

5

# Sample size estimation for cluster randomization designs

A quantitatively justified sample size calculation is almost universally regarded as a fundamental design feature of a properly controlled clinical trial. As stated by Friedman *et al.* (1996) 'clinical trials should have sufficient statistical power to detect differences between groups considered to be of clinical interest. Therefore calculation of sample size with provision for adequate levels of significance and power is an essential part of planning'. This statement would seem to apply with equal force to trials which randomize intact clusters as it does to trials which randomize individuals. It is particularly relevant to the design of community intervention trials, where many subjects and large expenditures of time and money have frequently been the norm. Yet, as discussed in Chapter 1, methodological reviews of cluster randomization trials have consistently shown that only a small proportion of these studies have adopted a predetermined sample size based on formal considerations of statistical power. It is interesting to speculate as to why this should be so. One obvious reason is that the appropriate formulas tend to be relatively inaccessible, not being given, for example, in most standard texts and articles on clinical trial methodology. A second reason is that the proper use of these formulas requires some prior assessment of the intracluster correlation coefficient $\rho$, either directly or through comparable information on the value of $\sigma_A^2$, the between-cluster component of variation. Neither of these parameters may be very familiar to investigators, complicating the task of obtaining relevant past data that may be used for sample size planning. This point was alluded to by Gail *et al.* (1992), in their discussion of the statistical aspects of the COMMIT trial, where they stated that 'the major problem in designing community intervention trials such as COMMIT is obtaining reliable estimates of the between-community component of variation'. A third reason why the issue of sample size estimation may tend to be ignored in cluster randomization trials is that studies enrolling hundreds or even thousands of patients may give the misleading impression of extensive statistical power, when in fact the effective sample size, after taking into account the clustering effect, is actually quite small.

Adding to the difficulties of ensuring that cluster randomization trials are of adequate size are several practical factors that more or less uniquely apply to the successful conduct of such studies. As will be seen, the power of a cluster randomization trial depends more on the number of units randomized than on their size. However, the number of units that can be realistically studied may be small, due to reasons of logistics and cost. Limited resources may also lead to a diluted intensity of effect in the

experimental group, esp
graphic area. These prol
subject blindness, thus
even assume, activities o
vention through a 'spill-c
follow-up due to unique

Many cluster random
adding to the difficulties.
lations, with the consequ
low in the first instance,
pliance with the interven
especially over a relativel
than usual risk in such s
necessarily those exposec

Published editorial co
context of community-
large-scale trials involvi
size of effects has been n
not even have the resou
similar point, goes on to

> outcome measures must b
> sized effect is meaningful
> make sure that we have t

These comments emphas
power is as important i
individuals.

In Section 5.1 we revie
size for cluster randomiz
adopted designs (i.e. con
described in Sections 5.2
up of study subjects and
provides some strategies
desired levels of statistica

## 5.1 General issu

There are a number of iss
domized trial. These inc
determination of a minin
of a statistical test or co
one- or two-sided). In thi
nificance in cluster rand
prior assessment of $\rho$.

As noted in Section 3.
interventions. For exam

experimental group, especially if these resources must be spread over a wide geographic area. These problems may be compounded by the difficulties of arranging subject blindness, thus allowing control group members to become aware of, or even assume, activities of the experimental group, further diluting the effect of intervention through a 'spill-over' effect. Finally there may be an increased risk of loss to follow-up due to unique features of this design.

Many cluster randomization trials take the form of prevention studies, further adding to the difficulties. Prevention trials usually enrol subjects from healthy populations, with the consequence that the event rates one wishes to reduce are relatively low in the first instance, with the benefits of intervention not readily apparent. Compliance with the intervention may also be difficult to maintain in such populations, especially over a relatively long period of time. More generally, there may be a greater than usual risk in such studies that the individuals evaluated for outcomes are not necessarily those exposed to the intervention.

Published editorial comments have explicitly dealt with these concerns in the context of community-based intervention trials. Susser (1995), commenting on large-scale trials involving public health interventions, noted that 'generally, the size of effects has been meagre in relation to the effort expended' and 'we often do not even have the resources to detect medium effects'. Fishbein (1996), making a similar point, goes on to emphasize that

> outcome measures must be sensitive to the purpose of the intervention, and when a small-sized effect is meaningful (and all we can expect) for a given outcome measure, we must make sure that we have the sample size necessary to detect such an effect.

These comments emphasize that the need to provide a formal estimate of statistical power is as important in trials randomizing clusters as it is in trials randomizing individuals.

In Section 5.1 we review some general issues involved in the estimation of sample size for cluster randomization trials. Methods applicable to the three most frequently adopted designs (i.e. completely randomized, matched-pair and stratified) are then described in Sections 5.2, 5.3 and 5.4 respectively. Issues involving loss to follow-up of study subjects and/or clusters are discussed in Section 5.5, while Section 5.6 provides some strategies which may help to overcome common barriers to achieving desired levels of statistical power.

## 5.1 General issues of sample size estimation

There are a number of issues common to sample size estimation that apply to any randomized trial. These include: (1) identification of the primary study outcome, (2) determination of a minimally important effect of intervention, and (3) specification of a statistical test or confidence interval method along with its directionality (i.e. one- or two-sided). In this section we focus on two related issues that have unique significance in cluster randomization trials: the determination of cluster size and the prior assessment of $\rho$.

As noted in Section 3.2, cluster size is to some degree, determined by the selected interventions. For example, households were the natural unit of randomization in

the study reported by Payment *et al.* (1991), which considered the effect of domestic water filters on subjects' risk of gastrointestinal disease. Consequently, the average cluster size at entry in this trial was approximately four. On the other hand, mass education studies such as COMMIT (COMMIT Research Group 1995a) must consider random assignment of entire communities. At times, however, investigators have much greater latitude in selecting the unit of randomization. For example, randomized trials examining the effect of vitamin A on childhood mortality have been designed allocating units as diverse as households, villages and entire districts to intervention groups.

It is important in practice to distinguish between the size of a cluster and the number of subjects sampled per cluster, i.e. subsampled. For instance, many community intervention trials are economically restricted to enrolling only a subset of eligible residents unless the primary outcome is based on routinely collected data (e.g. mortality). Thus in the COMMIT trial, approximately 550 heavy smokers were subsampled per community and then followed up to compare quit rates among subjects in the control and experimental groups.

Contamination between subjects in different intervention groups can be a problem whenever the clusters are geographically proximate. As a precaution against contamination in community intervention trials, Hayes (1998) suggested sampling subjects from the geographic center of each community. However, the degree of intracluster correlation is usually assumed, for the purposes of trial planning, to be unaffected by both the number of individuals subsampled and the method of subsampling used. Given the usual uncertainties associated with sample size estimation, this assumption should be reasonable in practice.

In some studies, the number of subjects sampled from each cluster may be determined so as to minimize the overall costs. This requires having some indication of the cost of enrolling an additional individual from each cluster relative to the cost of recruiting an additional cluster. Appropriate methods may be adapted using techniques developed for cluster sampling (e.g. Levy and Lemeshow 1980, Section 11.5, McKinlay 1994). However, it is usually simpler to determine the number of subjects sampled per cluster as part of a sensitivity analysis. These analyses are conducted by estimating the required trial size under a number of different scenarios. Investigators are often surprised to learn from such analyses that even small changes in the expected effect of intervention, in the number of subjects sampled per cluster or in the intracluster correlation coefficient can have large effects on the required sample size.

It may be argued that sensitivity analyses serve a particularly important role in the planning of cluster randomization trials. This is largely a consequence of the difficulty investigators have in obtaining accurate estimates of either between-cluster variability or intracluster correlation. Together with cluster size, these quantities are used to adjust the sample size for the variance inflation due to clustering (see Sections 5.2–5.4). However, inaccuracies may still remain because estimates of intracluster correlation obtained from studies with only a small number of clusters are very imprecise.

Difficulties in obtaining accurate estimates of intracluster correlation are further complicated by the relatively small number of publications which present these values when reporting trial results. However, intracluster correlation estimates for smoking-related outcomes as obtained from school-based intervention studies have been provided by Siddiqui *et al.* (1996) and Murray and Hannan (1990), while Slymen and Hovell (1997) provide similar outcomes obtained from a cluster randomized trial of 154 orthodontist offices. Estimated intracluster correlation coefficients and design effects for health-related outcomes as obtained from a worksite health promotion study have been provided by Kelder *et al.* (1993), while estimates of design effects for community trials of mortality from cardiovascular disease and cancer are given by Mickey *et al.* (1991) and Mickey and Goodwin (1993). Note that design effects always depend on the combined effect of the intracluster correlation coefficient and sizes of the clusters randomized.



An alternative approach to dealing with random variation in published estimates of intracluster correlation, based on conducting simulation experiments similar in principle to the bootstrap method, is described by Feng and Grizzle (1992). They propose that, instead of using a single value of $\rho$, one should simulate the results of studies of the size that yielded the estimate. One can then substitute the values calculated from each simulation into the appropriate formula to generate a distribution of these sample sizes, e.g. the 90th percentile, that reflects the degree of conservativeness desired. Yet another approach is to set the value of $\rho$ equal to the upper limit of a (say) 95 per cent confidence interval for this parameter as computed from a suitable dataset. One complication is the absence of strong evidence demonstrating that confidence intervals constructed using available variance estimators (e.g. Feng and Grizzle 1992, Murray *et al.* 1994) remain valid when these estimators are derived from studies involving only a few clusters of variable size. A further complication is that confidence interval methods for $\rho$ in the case of a binary outcome variable are not yet well developed. Setting $\rho = 1$ avoids the need to obtain a more accurate advance estimate of this parameter, but is clearly very conservative.

The degree to which responses of cluster members are correlated, and consequently the size of the resulting variance inflation factor, will tend to vary across different units of randomization. Not surprisingly, responses among subjects from smaller clusters (e.g. households) tend to be more highly correlated than responses among subjects from larger clusters (e.g. communities). For example, people from the same household tend to be more alike than randomly selected subjects who live in the same city. Other examples of this inverse relationship between cluster size and the degree of intracluster correlation have been noted repeatedly in both cluster randomization trials and surveys using cluster sampling (e.g. Hansen and Hurwitz 1942, Hansen *et al.* 1953, pp. 306–309, Katz *et al.* 1993b). Although the magnitude of the intracluster correlation coefficient tends to decline with cluster size, it does so at a relatively slow rate. Thus, larger variance inflation factors are usually obtained when randomizing large clusters such as communities than when randomizing much smaller clusters such as households.

Estimates of intracluster correlation may also be obtained from complex surveys which employ cluster sampling. For example, Gulliford *et al.* (1999) have reported a set of such estimates obtained from a cross-sectional survey of English adults which included data on a range of lifestyle risk factors and health outcomes. Verma and Le (1996) present a very extensive list of intracluster correlation coefficients for variables such as fertility, family planning, child health and mortality, as compiled from data generated by 48 nationally representative surveys. Summary information on these data sources is provided in Table 5.1.

**Table 5.1** Selected sources for estimates of intracluster correlation coefficients and design effects (Koepsell 1998)

| Reference | Outcomes | Cluster |
|---|---|---|
| Chen *et al.* (1997) | Cholesterol levels among children | Family |
| Vachon *et al.* (1998) | Dietary intakes among postmenopausal women | Family |
| Katz *et al.* (1993a, b), Katz and Zeger (1994) | Childhood illness | Family, village |
| Siddiqui *et al.* (1996) | Tobacco use among adolescents | Classroom, school |
| Murray and Hannan (1990), Murray *et al.* (1994) | Tobacco and drug use among adolescents | School |
| Slymen and Hovell (1997) | Tobacco and alcohol use among adolescents | Orthodontist offices |
| Kelder *et al.* (1993) | Physical health and tobacco use | Worksites |
| Gulliford *et al.* (1999) | Lifestyle risk factors and health outcomes | Neighbourhood, household |
| Verma and Le (1996) | Fertility rates, family and child health | Neighbourhood |
| Feldman and McKinlay (1994) | Height, weight, body mass index, blood pressure, cholesterol levels | Community |
| Hannan *et al.* (1994) | Behavioural risk factors, knowledge and attitudes concerning heart disease | Community |
| Murray and Short (1995) | Alcohol use among adolescents | Community |
| Murray and Short (1997) | Tobacco use among adolescents | Community |
| Mickey *et al.* (1991), Mickey and Goodwin (1993) | Mortality from cardiovascular disease and cancer | County |

To summarize, there are several reasons for considering a range of possible values of intracluster correlation when estimating sample size. First, estimates of intracluster correlation coefficient are dependent on the study design. Thus an estimate of intracluster correlation obtained from a study employing stratified randomization may be smaller than if a completely randomized design had been used (assuming that the stratification is effective). Second, this estimate may be computed from a study having only a small number of clusters. Third, the number of subjects participating per cluster in the planned trial will often be highly variable, which will further inflate the variance of the estimated effect of intervention. Finally, relatively little information is currently available concerning the generalizability of estimates of intracluster correlation. For example, the study referred to above by Verma and Le (1996) reports substantial variation in the degree of intracluster correlation across the 48 surveys examined.

## 5.2 The completely randomized design

### 5.2.1 Comparison of means

Suppose that $k$ clusters of $m$ individuals are randomly assigned to each of group $i$ ($i = 1, 2$), where $i = 1$ denotes the experimental group and $i = 2$ denotes the control group. We denote the primary response variable for an individual by $Y$, where $Y$ is assumed to be normally distributed with common but unknown variance $\sigma^2$. Since there are two sources of variation that influence the value of this parameter, we may write $\sigma^2 = \sigma_A^2 + \sigma_W^2$, where $\sigma_W^2$ denotes the variance in response within clusters and $\sigma_A^2$ the variance among clusters. We assume that the aim of the investigators is to test the hypothesis $H_0: \mu_1 = \mu_2$ at the two-sided $100\alpha$ per cent level of significance with power $1 - \beta$, where $\mu_1$ and $\mu_2$ are the population means of $Y$ in the experimental and control groups, respectively. Sample estimates of $\mu_1$ and $\mu_2$ are given by $\overline{Y}_1$ and $\overline{Y}_2$, respectively, where these estimates are computed over all individuals in each group.

Let $Z_{\alpha/2}$ denote the two-sided critical value of the standard normal distribution corresponding to the error rate $\alpha$, and $Z_\beta$ denote the critical value corresponding to $\beta$. Then, if $\overline{Y}_1 - \overline{Y}_2$ can be regarded as approximately normally distributed, the number of subjects required per intervention group is given (Donner *et al.* 1981) by

$$n = \frac{(Z_{\alpha/2} + Z_\beta)^2(2\sigma^2)[1 + (m-1)\rho]}{(\mu_1 - \mu_2)^2} \tag{5.1}$$

where $\rho = \sigma_A^2/(\sigma_A^2 + \sigma_W^2)$ is the intracluster correlation coefficient, and $\mu_1 - \mu_2$ denotes the magnitude of the difference to be detected (for one-sided calculations, $Z_{\alpha/2}$ is replaced by $Z_\alpha$). With this allocation, the 'effective sample size' for each group would be given by $n/[1 + (m-1)\rho]$. Thus at $\rho = 0$, equation (5.1) reduces to the usual sample size specification (e.g. Armitage and Berry 1994, Section 6.6). Equivalent to equation (5.1), the number of clusters required per group is given by

$$k = \frac{(Z_{\alpha/2} + Z_\beta)^2(2\sigma^2)[1 + (m-1)\rho]}{m(\mu_1 - \mu_2)^2} \tag{5.2}$$

The parameter $\mu_1 - \mu_2$ would usually be specified in advance as the minimum value of the intervention effect regarded as substantively important to detect. In practice, the investigators must assess this value on the basis of judgment, augmented by the best available data. Zucker *et al.* (1995) describe how this was done for the CATCH trial, where the investigators were required to specify the detectable mean difference for total serum cholesterol.

The use of the critical values $Z_{\alpha/2}, Z_\beta$ in the formulas above rather than critical values $t_{\alpha/2}$, $t_\beta$ corresponding to the $t$-distribution will underestimate the required sample size unless the degrees of freedom are large. A simple adjustment for this undercorrection (Snedecor and Cochran 1989, p. 104) is to add one cluster per intervention group when the sample size is determined using a 5 per cent type I error rate, and two clusters per group assuming a 1 per cent type I error rate. No correction for the degrees of freedom is necessary if the total number of clusters is approximately 30 or more (Lachin 1981).

A more exact iterative procedure may also be used, as described by Murray and Hannan (1990). Iteration is required to implement this procedure since the degrees of freedom, given by $2(k-1)$, are a function of sample size. However, these difficulties can be avoided using the SAS program provided by Donner and Klar (1996) to determine exact power.

In the case of unequal cluster sizes, we may replace $m$ in equation (5.1) by an advance estimate of the average cluster size $\overline{m}$. This approximation will tend to slightly underestimate the actual required sample size, but the underestimation will be negligible provided the variation in cluster size is not substantial. A conservative approach would be to replace $m$ by $m_{\max}$, the largest anticipated cluster size in the sample. Taking a conservative approach would also provide some protection

of statistical power in the event that the loss to follow-up rate for the trial is underestimated.

### Example 5.1

Hsieh (1988) reported on the results of a pilot study for a planned 5-year trial examining cardiovascular risk factors. Cholesterol levels (mg/dl) were obtained from 754 individuals from four worksites. The estimated value of the variance component within worksites is given by $S_W^2 = 2209$, while that between worksites is given by $S_A^2 = 93$. Therefore the value of the intracluster correlation coefficient may be assessed for the purpose of sample size estimation as

$$\rho = \frac{93}{(93 + 2209)} = \frac{93}{2302} = 0.04$$

Assuming approximately 70 eligible subjects per worksite, this implies that the variance inflation factor is given by $1 + (70 - 1)0.04 = 3.76$. The number of worksites which must be randomized to obtain 80 per cent power at $\alpha = 0.05$ (two-sided) for detecting a mean difference in cholesterol level of 20 mg/dl between intervention groups is then given by

$$k = \frac{(1.96 + 0.84)^2\, 2(2302)[1 + (70 - 1)0.04]}{70(20^2)} = 4.8$$

In practice, at least seven clusters might be randomly assigned per intervention group both to adjust for the use of critical values corresponding to the normal rather than the $t$-distribution and to allow for the possibility that one or two worksites might not participate for the full length of the trial. Then a total of 490 subjects would need to be recruited to each intervention group, assuming 70 subjects recruited per worksite and seven worksites per intervention group. Due to observed between-worksite variation in cholesterol levels, however, this sample size is effectively equivalent to an individually randomized trial with 130 subjects in each intervention group ($130 \simeq 490/3.76$). Any final determination of sample size should also consider a range of plausible values for both $\rho$ and $\sigma^2$, especially since the pilot study was limited to only four worksites. For example, if the true intracluster correlation $\rho = 0.10$, the variance inflation factor would be $1 + (70 - 1)0.10 = 7.9$, yielding $k = 10.2$.

### Further remarks

1. As noted by several authors (e.g. Koepsell *et al.* 1991, Murray and Short 1995) some improvement in power for cohort studies may almost always be obtained by taking into account the proportion of variation explained in the outcome measure by one or more baseline covariates. A natural choice for this purpose is the pre-test version of a post-test outcome measure.

   To account for the influence of a single covariate in the calculation of sample size, the value $(1 - r^2)\sigma^2$ could be substituted for $\sigma^2$ in equation (5.1), where $r$ is the anticipated correlation, possibly based on previous data, between the covariate and the outcome measure. For example, if $r = 0.5$, the required sample size may be reduced by 25 per cent, a saving which may be particularly attractive when resources are limited.



Adjustment for covariates measured at the cluster level (e.g. cluster size) should reduce the between-cluster source of variation and hence also reduce the value of $\rho$. In general this should also lead to an increase in power, depending on the relationship between the covariate and the outcome measure. This reflects the point made in Section 5.1, where, in considering the effect of stratification on the value of $\rho$, it was assumed that the strata are defined at the cluster level. However, the situation is more complicated when adjusting for baseline covariates measured at the individual level (e.g. age, sex). Increased power due to covariate adjustment may then be accompanied by either an increase or a decrease in the size of $\rho$. A more detailed discussion of the effect of covariate adjustment on the value of the intracluster correlation is provided by Stanish and Taylor (1983).

To account for the influence of several baseline covariates, the multiple correlation coefficient $R$ replaces $r$ in this expression, as illustrated by Murray and Short (1995). However, in many applications a suitable value of $R$ may not be available from empirical sources. In this case, it is useful to note that application of the relatively simple formula (5.1) will be conservative, i.e. it will tend to overestimate the actual required trial size.

Aside from increasing precision, recording a pre-test version of a post-test measure has the additional advantage of providing baseline data that can be used to reassess the values of $\sigma_A^2$, $\sigma_W^2$ and $\rho$, thus providing a check on the validity of sample size calculations that may rely heavily on the assumed values of these parameters.

2. In some trials, it may be considered more natural to frame the scientific question of interest in terms of a subject's change score from baseline rather than in terms of the final score alone. The primary analysis would then consist of comparing the average change in the experimental group with the corresponding average change in the control group. Assuming that the variation in the baseline and final scores is the same, the variance of a change score is given by $2\sigma^2(1 - r)$, where $r$ again denotes the correlation between the two scores. Then the appropriate size of sample is obtained by substituting $2\sigma^2(1 - r)$ for $\sigma^2$ in equation (5.1). This shows that it is statistically more efficient to compare the net changes in baseline between the two groups than it is to compare the final scores alone provided $r > 0.5$ (Fleiss 1986, Section 7.1). Note also that in this case the value of $\rho$ in equation (5.1) technically corresponds to the change score rather than the final score.
3. Although assignment of an equal number of clusters per intervention group is usually most efficient from a statistical perspective, practical considerations may occasionally suggest unequal allocation, as, for example, was done in the CATCH (Zucker *et al.* 1995). In this case, formula (5.2) may be easily adopted to calculate the required total sample size. Denoting the number of clusters to be enrolled in the experimental group by $k_1$, suppose we wish to enrol $k_2 = Qk_1$ clusters in the control group. Then to preserve the same error rate specifications as obtained under equal allocation with $k$ clusters per group ($Q = 1$), the numbers of clusters required under unequal allocation are given approximately by

$$k_1 = \tfrac{1}{2}k(1 + 1/Q) \qquad \text{and} \qquad k_2 = \tfrac{1}{2}k(1 + Q)$$

As would be expected, it can be easily shown that $K = k_1 + k_2$ exceeds $2k$, the total number of clusters required under equal allocation, where the increment