Number of Children Required in the Sample

In a survey many estimates may be made and for each variable, for a specified precision, there will be an appropriate sample size. Therefore, the most important estimate must be used as a basis for calculation of sample size. Further compromise is required between an ideal level of precision and the amount of time that a health department staff may be able to spend on field work.

The survey designs presented are oriented toward comparisons between two or more areas within a health jurisdiction and the calculation of sample sizes proceeds with this objective as a goal.

Testing of Hypotheses

Before entering into the details of calculation, some concepts of the Neyman-Pearson theory for testing of hypotheses will be briefly summarized, using the immunization survey as an example. As in the usual statistical test of significance, a "null hypothesis" of no significant difference between the proportion of adequately immunized 1-4 year-old children in each of the two areas, is set up. This will be written in abbreviated form as $H_0$.

The two possibilities for the true status of affairs, the null hypothesis, $H_0$, and the conclusions which may be drawn from sampling are shown in the following diagram.

<table>
<thead>
<tr>
<th>True Status of Proportions Immunized, $\pi_A$ and $\pi_B$ in Areas A and B.</th>
<th>Null Hypothesis $\pi_A = \pi_B$ in $H_0$</th>
<th>Sample Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi_A = \pi_B$</td>
<td>True</td>
<td>$H_0$ is Accepted $H_0$ is Rejected</td>
</tr>
<tr>
<td>$\pi_A \neq \pi_B$</td>
<td>False</td>
<td>Correct Conclusion Type I Error Probability $= \alpha$</td>
</tr>
</tbody>
</table>
The first column of the diagram lists two alternative true situations, \( \pi_A = \pi_B \) (immunization levels in the two areas are equal) and \( \pi_A \neq \pi_B \), (immunization levels in the two areas are not equal). The hypothesis to be tested, \( H_0 \), is stated in the heading of column (2). As shown, if \( \pi_A = \pi_B \), the hypothesis is true, if \( \pi_A \neq \pi_B \), the hypothesis is false. Columns (3) and (4) show the possible conclusions that may be drawn after examination of the sample. The investigator may decide either to reject the null hypothesis or to accept it. The rule by which the investigator decides whether to reject or accept is called a test of significance. If \( \pi_A = \pi_B \), as shown on the first line of column (1), the decision to reject the null hypothesis is incorrect and is called a Type I error. The probability, \( \alpha \), of making a Type I error, is established arbitrarily at the discretion of the investigator and, as shown below, is one of the factors used in calculation of sample size.

If \( \pi_A \neq \pi_B \), as shown on the second line of column (1), the decision to reject \( H_0 \), the null hypothesis, is correct. The probability, \( 1-\beta \), that \( H_0 \) will be rejected when the true immunization levels differ, depends on the magnitude of the difference between the true immunization levels, \( \pi_A \) and \( \pi_B \), as well as the sample size.

The fourth column of the diagram gives a similar analysis of the decision to accept the null hypothesis. If, \( \pi_A = \pi_B \), as shown on the first line of column (1), the conclusion drawn from the sample findings is correct and will be made with probability \( 1-\alpha \). If, as on the second line, \( \pi_A \neq \pi_B \), the decision to accept the null hypothesis is incorrect and is called a Type II error. \( \beta \) designates the probability of making a Type II error.

**The Type I Error**

In practice the sample proportions \( p_A \) and \( p_B \) will differ, even when \( \pi_A = \pi_B \), because of sample variation. Therefore, a range of differences, \( p_A - p_B \), which may be attributed to sampling variation must be established. The upper and lower limits of this range are defined as critical levels, which will be denoted as \( \pm h \). Sample differences with absolute values greater than \( h \) are said to fall in the critical region. Thus \( \alpha \) is the
probability, (Figure 15) when \( \pi_A = \pi_B \), that the difference between two sample proportions will fall in the critical region.

**Figure 15. Probability \( \alpha \), of a Type I Error**

In the method of this chapter sample sizes will be calculated with specification of the probability of making a Type I error. The probability of making a Type II error is then evaluated for each sample size and specified values of \( \pi_1 \) and \( \pi_2 \). The parameter which is actually calculated is \( 1 - \beta \), the probability that a true difference will be detected.

*Calculation of \( m \), the Required Number of Children*

Two assumptions are made in the following calculations: (1) probabilities may be approximated with reasonable accuracy by the normal distribution and (2) the variance of the difference between the two sample proportions may be estimated by

\[
\frac{\sigma^2}{p_A - p_B} = \text{Est. var} (p_A - p_B)
\]

\[
= \frac{g}{m} [\pi_A (1 - \pi_A) + \pi_B (1 - \pi_B)].
\]  

This is the usual formula for the variance of a difference between two proportions multiplied by a "clustering coefficient," \( g \), to compensate for increase in variance resulting from within-family correlation in immunization status of children, and to a lesser extent
from correlation between families within primary sampling units.

With these assumptions the formula for sample size is developed from the equation

\[
h = \frac{z_{\alpha/2}}{\hat{\sigma}} \sqrt{p_A - p_B}.
\]  

(11.8)

In which,

\[
h = \text{the critical level, assigned arbitrarily. Sample sizes will be calculated for a range of values of } h.\]

\[
z_{\alpha/2} = \text{the standardized deviation demarking an area equal to } \alpha/2 \text{ on each tail of the normal probability distribution. The magnitude of } \alpha \text{ is set by the investigator. In the present calculations } \alpha/2 = .025 \text{ and the corresponding value of } z_{\alpha/2} \text{ is } 1.96.\]

For given values of \( g \) and \( m \), formula (11.7) will have its maximum value when \( \pi_A = \pi_B = 1/2 \). Making this substitution for \( \pi_A \) and \( \pi_B \),

\[
\hat{\sigma}^2_{p_A - p_B} = g \frac{2}{2m}.
\]  

(11.9)

Substituting the square root of (11.9) for \( \hat{\sigma}^2_{p_A - p_B} \) in (11.8) gives

\[
h = \frac{z_{\alpha/2}}{\frac{g}{2m}} \left[ \frac{g}{2m} \right]^{1/2},
\]  

(11.10)

and this equation, solved for \( m \), is

\[
m = \frac{g}{2} \left[ \frac{z_{\alpha/2}}{h} \right]^2.
\]  

(11.11)

An estimate of the value of \( g \) was obtained from the regression

\[
\hat{\sigma}^2_s = g \hat{\sigma}^2_b,
\]  

(11.12)
in which $\sigma_s^2$ is the sector variance computed by Formula (9.1) and $\sigma_b^2$ the binomial variance computed by (11.4). Using data for the proportion of immunized 1-4 year-old children in several survey areas, the values of $g$ shown in Table 34, were obtained.

Table 34. Estimates of $g$, the Clustering Coefficient

<table>
<thead>
<tr>
<th>Immunization</th>
<th>Age Groups</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1-4</td>
<td>5-14</td>
<td>15-39</td>
</tr>
<tr>
<td>Oral Poliovaccine</td>
<td>1.69</td>
<td>2.61</td>
<td>1.81</td>
</tr>
<tr>
<td>Inactivated Poliovaccine</td>
<td>1.51</td>
<td>3.06</td>
<td>1.54</td>
</tr>
<tr>
<td>Diphtheria-Pertussis-Tetanus</td>
<td>1.32*</td>
<td>2.62**</td>
<td>***</td>
</tr>
<tr>
<td>Smallpox (ever)</td>
<td>1.85</td>
<td>2.03</td>
<td>1.51</td>
</tr>
<tr>
<td>Average</td>
<td>1.59</td>
<td>2.58</td>
<td>1.62</td>
</tr>
</tbody>
</table>

*Primary Series, **Booster in last 4 years, ***Data not collected.

Substituting the values $g = 1.6$ and $z_{\alpha/2} = 1.96$ in Formula (11.11) gives the working formula,

$$m = \frac{3.0733}{h^2},$$  \hspace{1cm} (11.13)

from which the values of $m$ in Table 4, (page 24) were calculated for the indicated values of $h$.

Calculation of $1 - \beta$

In this calculation values of $\pi_A$ and $\pi_B$ must be specified. The notation will be changed to $\pi_L$, representing the proportion immunized in a low socioeconomic area or in any area before carrying out an immunization program and $\pi_U$, representing the proportion immunized in an upper socioeconomic area, or in any area at the close of an immunization program. Thus $\pi_U$ will always represent an area with a higher and $\pi_L$ an area with a lower immunization level. The true difference in immunization rates will be represented by

$$\pi_U - \pi_L = \kappa.$$  \hspace{1cm} (11.14)
These relationships are illustrated in Figure 16.

Figure 16. Probability that Sample Differences Exceed the Critical Level when \( \pi_U \) is greater than \( \pi_L \)

From Figure 16,

\[ h - \kappa = z_{1-\beta} \frac{\sigma p_U - p_L}{p_U - p_L}, \]

from which

\[ z_{1-\beta} = \frac{h - \kappa}{\sigma p_U - p_L}. \]

(11.15)

The critical level, \( h \), will usually be smaller than \( \kappa \) as shown in Figure 16. In this case \( h - \kappa \) will be negative and consequently \( z_{1-\beta} \) will also be negative. The probability \( 1-\beta \) will then be equal to the area under the normal curve from \( z_{1-\beta} \) a negative quantity, to \( z = +\infty \).

Variance formula (11.7) has an absolute maximum when \( p_U = p_L = 1/2 \), but for values of \( p_U \) or \( p_L \neq 1/2 \), it will have a relative maximum when

\[ p_U + p_L = 1, \]

(11.16)
a condition for which \( 1-\beta \) will be a minimum for given values of \( h \), \( \kappa \), and \( m \).
Equations (11.14) and (11.16) may be solved simultaneously, yielding

\[ \pi_U = \frac{1+\kappa}{2}, \quad \pi_L = \frac{1-\kappa}{2}, \]

and if these are substituted in equation (11.7) it becomes

\[ \frac{\sigma^2}{p_U - p_L} = \frac{g}{2m} (1-\kappa^2). \]  

(11.17)

With substitution of (11.17) into (11.15)

\[ z_{1-\beta} = (h-\kappa) \sqrt{\frac{2m}{g(1-\kappa^2)}}, \text{ which} \]  

(11.18)

with \( g = 1.6 \), gives the working formula

\[ z_{1-\beta} = (h-\kappa) \sqrt{\frac{1.25m}{1-\kappa^2}}. \]  

(11.19)

The values of \( 1-\beta \) in Table 4 (page 24) were obtained by substitution of \( m, h \) and \( \kappa \) in equation (11.19). For example, with

\[ m = 50 \]
\[ h = 0.25 \]
\[ \kappa = 0.40 \]

\[ z_{1-\beta} = -1.15 \sqrt{\frac{1.25(50)}{0.84}} = -1.15(8.626) = -1.29. \]

\( 1-\beta \), the area under the normal curve to the right of \( z = -1.29 \), equals 0.90. With \( m = 50 \) and a critical level of \( h = 0.25 \) for \( \alpha = 0.05 \), the probability of detecting a true difference of 40 percent in immunization levels is at least 0.90.

This result shows that if samples of 50 children be drawn from each of two populations for which the true difference in the proportion immunized is 0.40, it will be found that in 90 percent or more of such pairs of samples, the difference between the sample proportions will exceed a critical level established at 0.25.