

## 1:1 Matched Case-Control Studies

Let  $p_1$  and  $p_0$  denote the probabilities of exposure for cases and controls, respectively.

The probability of a discordant pair is  $p_1(1 - p_0) + p_0(1 - p_1)$ .

If we require  $T$  discordant pairs to achieve the desired power, then the total number of pairs needs to be

$$N = \frac{T}{p_1(1 - p_0) + p_0(1 - p_1)}.$$

If  $p_1$  and  $p_0$  are small, this could be a large number.

How to estimate  $T$ ?

The required sample size is the one for which we will reject the null hypothesis with the desired probability, under a specified alternative hypothesis.

Here we assume that the unexposed group rate is  $P_0$  and the alternative is represented by  $\psi = \psi_A$ .

To compute the probability of rejection, we first need the form of the test of the null hypothesis.

The null hypothesis  $H_0 : \psi = 1$  is equivalent to

$$H_0 : \pi = \frac{\psi}{1 + \psi} = \frac{1}{2}.$$

McNemar's test (without continuity correction) for  $H_0$  is

$$\frac{(n_{01} - n_{10})^2}{n_{01} + n_{10}} = \frac{(2n_{10} - T)^2}{T} \sim \chi_1^2.$$

Taking square root,

$$Z = \frac{2n_{10} - T}{\sqrt{T}} \sim N(0, 1)$$

under  $H_0$  asymptotically.

We reject  $H_0$  if  $|Z| > z_{\alpha/2}$  where  $z_{\alpha/2}$  is the upper  $\alpha/2$  quantile of the standard normal distribution.

Under an alternative hypothesis ( $\psi = \psi_A$ ),  $n_{01} \sim \text{Bin}(T, \pi_A)$  given  $T$  where

$$\pi_A = \frac{\psi_A}{1 + \psi_A}.$$

Therefore,

$$E[Z|\pi_A] = \frac{2\pi_A T - T}{\sqrt{T}} = (2\pi_A - 1)\sqrt{T}$$

and

$$\text{Var}(Z|\pi_A) = \frac{4\pi_A(1 - \pi_A)T}{T} = 4\pi_A(1 - \pi_A).$$

The upper half of the rejection region is  $Z > z_{\alpha/2}$  which happens if and only if

$$\frac{Z - (2\pi - 1)\sqrt{T}}{\sqrt{4\pi(1 - \pi)}} > \frac{z_{\alpha/2} - (2\pi - 1)\sqrt{T}}{\sqrt{4\pi(1 - \pi)}}$$

and, asymptotically,

$$\frac{Z - (2\pi - 1)\sqrt{T}}{\sqrt{4\pi(1 - \pi)}} \sim N(0, 1).$$

So we have power  $1 - \beta$  if  $\Pr\{Z > z_{\alpha/2} | \pi\} = 1 - \beta$  which happens if and only if

$$\frac{z_{\alpha/2} - (2\pi - 1)\sqrt{T}}{\sqrt{4\pi(1 - \pi)}} > -z_{\beta}.$$

If we solve for  $T$ ,

$$T = \left( \frac{z_{\alpha/2} + 2\sqrt{\pi(1 - \pi)}z_{\beta}}{2\pi - 1} \right)^2.$$

Example:

$p_0 = 0.1$ ,  $\psi_A = 3$ , therefore  $\pi_A = 3/4 = 0.75$ , and

$$p_1 = \frac{0.1 \times 3}{1 - 0.1 + 0.1 \times 3} = 0.25.$$

Therefore,

$$N = \frac{T}{0.1 \times 0.75 + 0.9 \times 0.25} = \frac{T}{0.3}.$$

At  $\alpha = 0.05$ , and  $1 - \beta = 0.9$ ,  $z_{\alpha/2} = 1.96$ ,  $z_{\beta} = 1.28$ .

$$T = \left( \frac{1.96 + 2\sqrt{0.75 \times 0.25} \times 1.28}{2 \times 0.75 - 1} \right)^2 = 37.67$$

and

$$N = \frac{37.67}{0.3} = 125.5 \text{ pairs.}$$

## Non-Matched Case-Control Studies

Consider that the the data is summarized as a  $2 \times 2$  table and that we have sampled from the rows.

Suppose that  $n_1 = n_0$  and let  $\bar{p} = (p_0 + p_1)/2$ .

Then

$$Z = \frac{(a - n_1 m_1 / N) N^{3/2}}{\sqrt{n_1 n_0 m_1 m_0}}.$$

Asymptotically,  $Z \sim N(0, 1)$  under  $H_0$ , and

$$\begin{aligned} E[Z] &= \frac{(n_1 p_1 - n_1 (p_1 n_1 + p_0 n_0) / N) N^{3/2}}{\sqrt{n_1 n_0 (p_1 n_1 + p_0 n_0) (N - p_1 n_1 + p_0 n_0)}} \\ &= \frac{(p_1 - p_0) \sqrt{N}}{2 \sqrt{\bar{p} (1 - \bar{p})}} \end{aligned}$$

and

$$\text{Var}(Z) = \frac{p_1(1 - p_1) + p_0(1 - p_0)}{2\bar{p}(1 - \bar{p})}$$

under  $H_A$ .

By Jensen's inequality,  $\text{Var}(Z) < 1$  if  $p_1 \neq p_0$ .

Using a similar computation to that above, we have

$$N = 2 \left( \frac{z_{\alpha/2} \sqrt{2\bar{p}(1-\bar{p})} + z_{\beta} \sqrt{p_0(1-p_0) + p_1(1-p_1)}}{p_1 - p_0} \right)^2.$$

By substituting

$$p_1 = \frac{\psi_A p_0}{1 - p_0 + \psi_A p_0},$$

we may write the sample size in terms of  $p_0$  and  $\psi$ .

If  $p_0$  is small, then  $p_1 \approx \psi_A p_0$ .

Example:

Suppose  $p_0 = 0.1$ ,  $\psi_A = 3$ ,  $p_1 = 0.25$  and thus  $\bar{p} = 0.175$ .

Then, with  $\alpha = 0.05$  and  $1 - \beta = 0.9$ ,

$$\begin{aligned} N &= 2 \left( \frac{1.96 \sqrt{2 \times 0.175(1 - 0.175)} + 1.28 \sqrt{0.25 \times 0.75 + 0.1 \times 0.9}}{0.25 - 0.1} \right)^2 \\ &= 265 \text{ (total subjects)}. \end{aligned}$$

Note that the matched study requires 252 total subjects.

## Cohort Studies: Comparison with an External Standard

$$\text{SMR} = \frac{O}{E}$$

With  $E$  considered fixed,

$$O \sim \text{Poisson}(\theta E).$$

First, estimate of  $E$  is required.

Then, determine cohort size to yield  $E$  if possible.

$$\sqrt{O} \sim N(\sqrt{\theta E}, \frac{1}{4})$$

$$\sqrt{E}(\sqrt{\theta} - 1) = \frac{z_{\alpha/2}}{2} + \frac{z_{\beta}}{2}$$

Thus

$$E = \frac{(z_{\alpha/2} + z_{\beta})^2}{4(\sqrt{\theta} - 1)^2}.$$

Example:  $\alpha = 0.05$ ,  $1 - \beta = 0.8$

$E = 39$  when  $\theta = 1.5$

$E = 12$  when  $\theta = 2$

Finding cohort size to yield  $E = 12$  requires knowledge of standard rates, age distribution of cohort, etc.

Cohort Studies: Comparison within Cohorts  
 2 subgroups, "exposed (1)" and "unexposed (2)"

$$O_+ = O_1 + O_2$$

Assume same size and age distributions.

Given  $O_+$ ,

$$O_1 \sim \begin{cases} \text{Bin}(O_+, \frac{1}{2}), & \text{under } H_0 \\ \text{Bin}(O_+, \frac{\theta}{\theta+1}), & \text{under } H_1 \end{cases} .$$

Asymptotically,

$$O_1 \sim \begin{cases} N(\frac{1}{2}O_+, \frac{1}{4}O_+), & \text{under } H_0 \\ N(\frac{\theta}{\theta+1}O_+, \frac{\theta}{(\theta+1)^2}O_+), & \text{under } H_1 \end{cases} .$$

$$O_+ \left( \frac{\theta}{\theta+1} - \frac{1}{2} \right) = z_{\alpha/2} \sqrt{\frac{O_+}{4}} + z_{\beta} \sqrt{\frac{O_+ \theta}{(\theta+1)^2}}$$

Thus

$$O_+ = \frac{\left( \frac{z_{\alpha/2}}{2} + \frac{z_{\beta} \sqrt{\theta}}{\theta+1} \right)^2}{\left( \frac{\theta}{\theta+1} - \frac{1}{2} \right)^2} .$$

Example:  $\alpha = 0.05$ ,  $1 - \beta = 0.8$

$O_+ = 194$  when  $\theta = 1.5$ , and thus  $O_2 = 78$

$O_+ = 68$  when  $\theta = 2$ , and thus  $O_2 = 23$

About twice as many as before times 2 groups, i.e. 4 times as many

Compare this to  $\text{Var}(\bar{X} - \mu) = \sigma^2/n$  (for one sample problem) and  $\text{Var}(\bar{X}_1 - \bar{X}_2) = 2\sigma^2/n$  (for two sample problem), where  $n$  is size of a group.

Alternatively,

$$O_+ = \frac{(z_{\alpha/2} + z_{\beta})^2}{4 \left\{ \sin^{-1} \sqrt{\frac{\theta}{\theta+1}} - \sin^{-1} \frac{1}{2} \right\}^2}.$$

If exposed to unexposed ratio is 1:k, then

$$O_+ = \frac{(z_{\alpha/2} + z_{\beta})^2}{4 \left\{ \sin^{-1} \sqrt{\frac{\theta}{\theta+k}} - \sin^{-1} \frac{1}{1+k} \right\}^2}.$$

## General Considerations based on Likelihood Theory

Note

$$\hat{\theta} \sim N(\theta, I^{-1}(\theta))$$

where

$$I(\theta)_{i,j} = -E \left[ \frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_j} \right]$$

and  $l(\theta)$  is the log likelihood function.

Also an overall test of  $H_0 : \theta = \theta_0$  is given by

$$(\hat{\theta} - \theta_0)^T I(\theta_0) (\hat{\theta} - \theta_0) \sim \chi_p^2.$$

Power and sample size considerations can then be approached through the distribution of the quadratic form above under the specified alternative with  $\theta_1$ :

$$N = \frac{\{z_{\alpha/2} i^{-1/2}(\theta_0) + z_{\beta} i^{-1/2}(\theta_1)\}^2}{(\theta_1 - \theta_0)^2}$$

where  $i(\theta)$  refers to the expected information in a single observation.