9.2. How is equation (9.2) derived?

In order to have $1-\beta$ power and a two-sided $\alpha$ significance level to detect a difference $\pi_2-\pi_1$ between two groups (or in this case surveys), with group 1 of size $N_1$ and group 2 of size $N_2$, equation (1) needs to hold.

$$\begin{align*}
(z_{\alpha/2} + z_{\beta})^2 \times \left\{ \frac{\pi_1(1-\pi_1) + (m_1-1) \pi_1^2 k_1^2}{N_1} + \frac{\pi_2(1-\pi_2) + (m_2-1) \pi_2^2 k_2^2}{N_2} \right\} = (\pi_2-\pi_1)^2 \tag{1}\end{align*}$$

The formula allows cluster size $m$ (denoted in equation (1) as $m_1$ and $m_2$ for the first and second surveys respectively) to vary between the two surveys. It also allows for the sample size of the first survey $N_1$ to be different from the sample size $N_2$ of the second survey.

Equation (1) is based on first principles of power calculation.

- In the special case of repeat surveys, $N_i$ denotes the size of the first survey, is fixed, and cannot be changed. The most logical prior guess to use for replacing $\pi_i$ is $p_i$ (i.e. the point estimate from the first survey). Also, the most logical prior guess to use for $k_i$ is fixed, as it is calculated from the first survey. For consistency with equation 9.1 in the book, we denote $\pi_2=\pi$ and $N_2=N$.

- Additionally, in the special case of a repeat survey, we can use a one-sided significance level $\alpha$, since we are mostly interested in having power to show a fall in prevalence rather than being equally interested in demonstrating statistical evidence of an increase. Therefore, we can change $z_{\alpha/2}$ to $z_{\alpha}$ i.e. from 1.96 to 1.65 in the case of a 5% significance level.

$$\begin{align*}
(1.65+z_{\beta})^2 \times \left\{ \frac{p_1(1-p_1) + (m_1-1) p_1^2 k_1^2}{N_1} + \frac{\pi(1-\pi) + (m_2-1) \pi^2 k_2^2}{N} \right\} = (\pi-p_1)^2 \tag{2}\end{align*}$$

- Equation (2) does not have a solution (for $N$) for all combinations of $N_i$, $k_i$, $m_i$, $p_i$, and $\pi$. This makes sense, because to detect a difference between two surveys a minimum sample size in each of the two surveys is required, therefore for a given $N_i$ (sample size of the first survey) there is a minimum change in prevalence for which there is power to detect it. That is, however big the second survey is, sometimes the first survey is just too small to detect a small change in prevalence.

We can solve for $N$ by rearranging equation (2):

$$N = \frac{(1.65+z_{\beta})^2 \times \{\pi(1-\pi) + (m_2-1) \pi^2 k_2^2\} \times N_1}{\pi(1-p_1)^2 \times N_i -(1.65+z_{\beta})^2 \times \{p_1(1-p_1) + (m_1-1) p_1^2 k_1^2\}} \tag{3}$$

**Example 1:**

Following on from the example 9.2 in the book with a one-sided $\alpha$ ($z_{\alpha}=1.65$), 80% power ($z_{\beta}=0.84$), $\pi_1=0.00366$, $\pi_2=0.00256$, $m_1=528$ and $m_2=600$, $k_1=0.31$, $k_2=0.54$ and $N_i$ set at 22000, we cannot solve equation (2) as the denominator becomes negative.

However, if we wanted to detect a 35% fall in prevalence from $\pi_1=366$ per 100,000 to $\pi_2=238$ per 100,000, then solving for $N=49720$. This sample size would still need to be inflated to allow for non-participation.