Sample Size Caluclations

Casey Breen

In this notebook, I calculate the sample size required to detect a minimum detectable effect (MDE) in the mortality rate of 0.4 (per day / 10,000 people). The minimum detectable effect (MDE) is the effect size above which we can statistically distinguish our estimate from 0.

Minimum Detectable Effect

We can calculate the minimum detectable effect as:

$$\begin{aligned} MDE &= se(\hat{\delta})[\Phi^{-1}(1-\alpha/2) + \Phi^{-1}(1-\beta)] \\ &= \sqrt{p\left(\frac{\text{deff}_1}{n_1\bar{d_1}} + \frac{\text{deff}_2}{n_2\bar{d_2}} + \frac{\lambda}{p + n_1 + \bar{d_1}}\right)}[\Phi^{-1}(1-\alpha/2) + \Phi^{-1}(1-\beta)] \end{aligned}$$

where

- p is the mortality rate per day per 10,000 people
- n_1 is the network survival sample size
- n_2 is the number of households reported
- $\bar{d_1}$ is the average degree
- $\bar{d_2}$ is the average household size
- deff_1 is the design effect for the network survival survey
- deff_2 is the design effect for the household survey
- α is our confidence level (generally 0.05)
- λ is the true difference between the household survey estimand and the non-probability estimand
- $1 - \beta$ is our power (generally 0.8 or 0.9)

Sample Size Calculations

For convenience, make several assumptions to estimate our required sample size:

- $\alpha = 0.05$ (confidence level)
- $1 - \beta = 0.9$ (power)
- $p = 0.87$ deaths per day per 10,000 people

- $\text{deff}_1 = \text{deff}_2 = 1.5$ (design effect)
- $\bar{d_1} = 20$ (personal network size)
- $\bar{d_2} = 5.5$ (household size)
- $n_2 = 1{,}074$ (household sample
- n_1 - systematically varied from 50 to 1,000, by intervals of 50

Mortality rate adjustment

Before applying the above formula, we need to convert our estimates of mortality rates into proportions. For example, we can convert the following mortality (MR):

$$\text{Mortality Rate (MR)} = \left(\frac{\text{Total Deaths During the Period}}{\text{Num. of people observed} \times \text{Num. of observed days}} \right) \times 10,000$$

into a proportion p using the following equation:

$$\begin{aligned} P &= \left(\frac{\text{Num. Total Deaths During the Period}}{\text{Num. of people observed}} \right) \\ &= MR \times \frac{\text{Num. of observed days}}{10,000} \end{aligned}$$

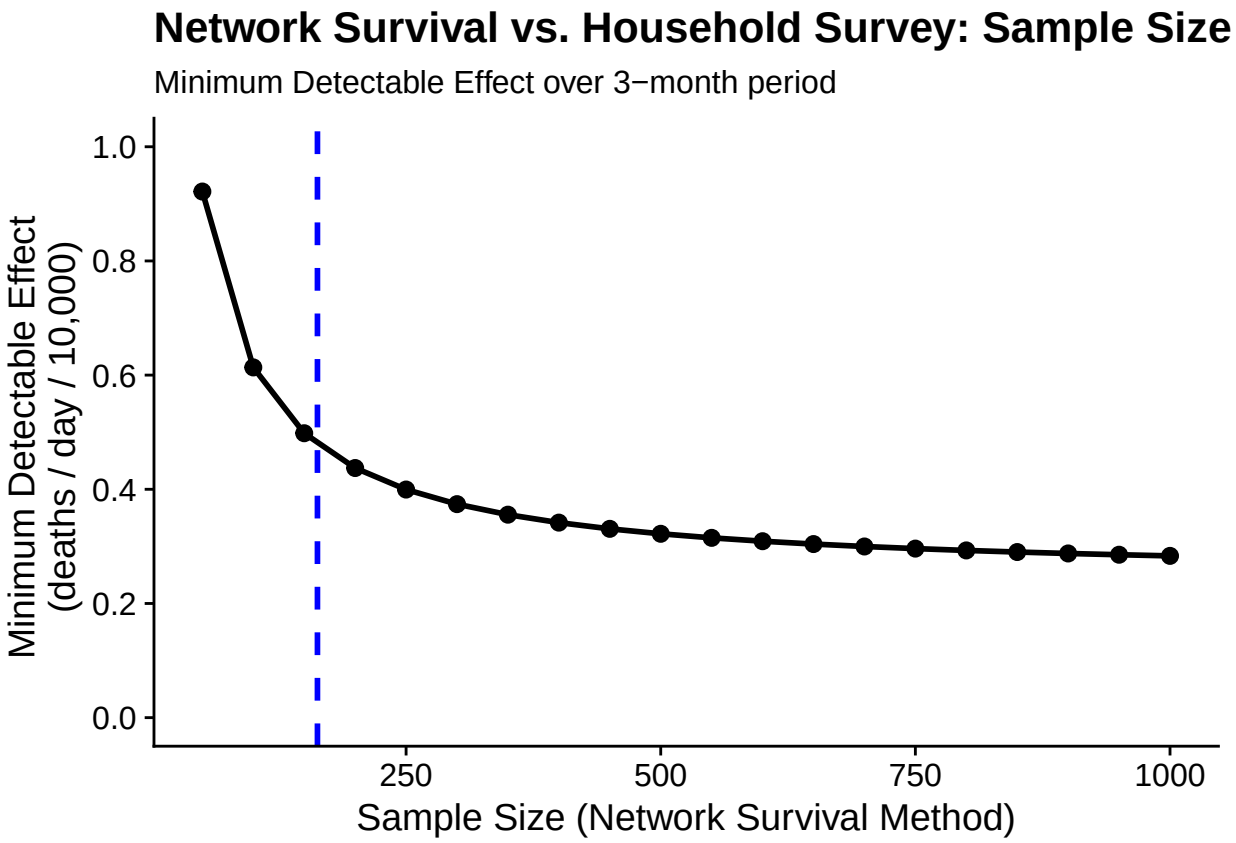
This allows us to use the MDE equation above to estimate mortality. Here, we assume that the different length of reporting windows (e.g., 60 days vs. 180 days) are accounted for when we transform the mortality rate into a proportion of individuals who are during the reporting period.

Sample Size Calculation – Household vs. Network Survival

In one health zone, what is the sample size required to detect a minimum difference of 0.4 in the estimated mortality rate (in deaths / day / 10,000) between our household and the network survival study, over a 3 month observation window?

```
## library
library(tidyverse)
library(data.table)

## assumptions
alpha = 0.975                                  ## Precision
beta = 0.1                                     ## Power = 1 - beta
nsm_sample <- seq(50, 1000, by = 50)           ## network survival method sample size
nsm_de <- 1.5                                  ## network survival method design effect
hh_de <- 1.5                                   ## household design effect
nsm_network <- 30                              ## network survival method
mrate <- 0.87                                  ## mortality rate (deaths / day / 10,000)
time_window <- 90                              ## time window
lambda <- 0.4                                  ## difference between HH and NSM surveys
lambda_converted <- lambda * time_window/10000 ## convert lambda to correct scale
mr_porportion <- mrate * time_window/10000     ## proportion of sample dying
```

```r
## estimate mde
## assume for simplicity difference is 0.4
mde <- sqrt(mr_porportion * ((nsm_de/(nsm_network * nsm_sample)) + (hh_de/(1074*5.5)) +
       0.4/(mr_porportion*nsm_sample*nsm_network)) * (qnorm(0.975) * qnorm( 1-beta)))

## recalculate with more exact lambda values (mde)
mde <- sqrt(mr_porportion * ((nsm_de/(nsm_network * nsm_sample)) + (hh_de/(1074*5.5))
       + mde/(mr_porportion*nsm_sample*nsm_network)) * (qnorm(0.975) * qnorm( 1-beta)))

## convert back to deaths/day/10000
mde_convert <- mde * 10000/time_window

## visualize
tibble(nsm_sample, mde_convert) %>%
  ggplot(aes(x = nsm_sample, y = mde_convert)) +
  geom_point(size = 2.5) +
  geom_line(size = 1) +
  scale_y_continuous(limits = c(0, 1), n.breaks = 8) +
  geom_vline(xintercept = 163, linetype = "dashed", size = 1, color = "blue") +
  labs(x = "Sample Size (Network Survival Method)",
       y = "Minimum Detectable Effect \n (deaths / day / 10,000)",
       title = "Network Survival vs. Household Survey: Sample Size",
       subtitle = "Minimum Detectable Effect over 3-month period") +
  cowplot::theme_cowplot()
```



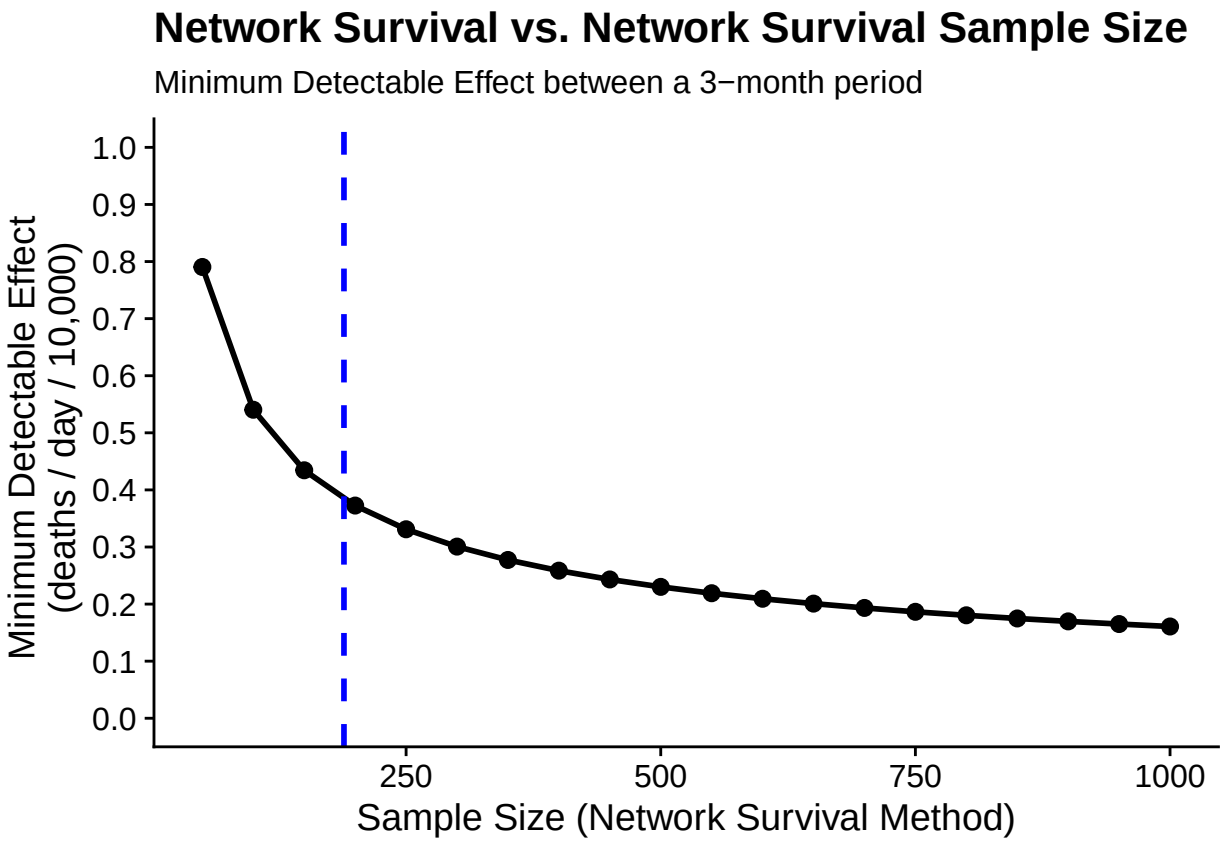
We will need a sample size of of approximately N = 163 to detect a minimum effect of 0.4 over 3 month period between the network survival estimate and the household survey estimate (N = 1,074, household size

= 5.5).

Sample Size Calculation – Network Survival vs. Network Survival

In one health zone, what is the sample size required to detect a minimum effect of 0.4 for the estimated mortality rate (in deaths / day / 10,000) between to network survival method samples?

```
## assumptions
alpha <- 0.975                                  ## Precision
beta <- 0.1                                     ## Power = 1 - beta
nsm_sample <- seq(50, 1000, by = 50)            ## network survival method sample size
nsm_de <- 1.5                                   ## network survival method design effect
nsm_network <- 30                               ## network survival method
mrate <- 0.87                                   ## mortality rate (deaths / day / 10,000)
time_window <- 90                               ## time window
mr_porportion <- mrate * time_window/10000      ## proportion of sample dying

## estimate mde
mde <- sqrt(mr_porportion* ((nsm_de/(nsm_network * nsm_sample)) + (nsm_de/(nsm_network * nsm_sample)) +

## recalculate with more exact lambda values (mde)
mde <- sqrt(mr_porportion* ((nsm_de/(nsm_network * nsm_sample)) + (nsm_de/(nsm_network * nsm_sample)) +

## convert back to deaths/day/10000
mde_convert <- mde * 10000/time_window

## visualize
tibble(nsm_sample, mde_convert) %>%
  ggplot(aes(x = nsm_sample, y = mde_convert)) +
  geom_point(size = 2.5) +
  geom_line(size = 1) +
  scale_y_continuous(limits = c(0, 1), n.breaks = 10 ) +
  geom_vline(xintercept = 189, linetype = "dashed", size = 1, color = "blue") +
  labs(x = "Sample Size (Network Survival Method)",
       y = "Minimum Detectable Effect \n (deaths / day / 10,000)",
       title = "Network Survival vs. Network Survival Sample Size",
       subtitle = "Minimum Detectable Effect between a 3-month period") +
  cowplot::theme_cowplot()
```



We will need a sample size of N = 189 for the network survival method (per period) to detect a minimum effect size of 0.4 over two consecutive 3-month periods.