EMD538 Problem Set 1

Cole Brookson

18 Semptember, 2024

Consider the data presented by Greenwood (1931) on the distribution of secondary measles cases among households with 1 index case and m = 4 contacts.

Question 1

What is the secondary attack rate (SAR) for the households with $m = 4$ contacts, assuming all infected contacts had symptom onset within the maximum serial interval from the primary case?

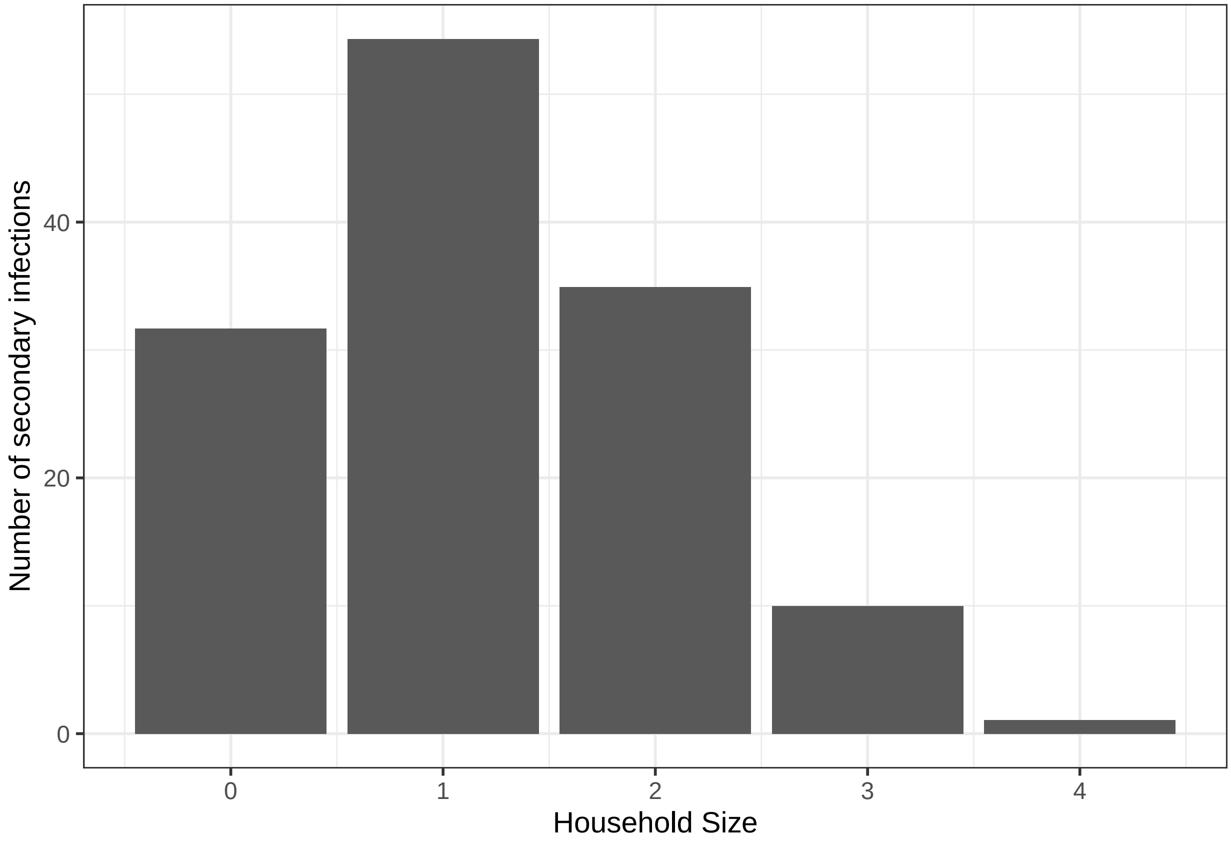
```
numerator <- sum(infect * infect_freq[, 4]) # number of people infected
denominator <- sum(infect_freq[, 4] * 4) # number of people susceptible
p <- numerator / denominator
print(paste0("The SAR for households with m = 4 is ", p))
```

```
## [1] "The SAR for households with m = 4 is 0.265151515151515"
```

Question 2

What is the expected distribution of secondary infections within households of size $m = 4$ if there is NO ongoing transmission and $p = 0.3$?

```
df_gw <- data.frame(
    sec_inf = sum(infect_freq[, 4]) * dbinom(0:4, 4, 0.3),
    hhs = c(0:4)
)
ggplot2::ggplot(df_gw) +
    ggplot2::geom_col(ggplot2::aes(x = hhs, y = sec_inf)) +
    ggplot2::labs(x = "Household Size", y = "Number of secondary infections") +
    ggplot2::theme_bw()
```



Question 3

What are the possible chains that could result in all 4 contacts being infected? What is the probability of each chain (in terms of p and q) under the Greenwood assumption?

If p is the probability of infection and $q = 1 - p$, there are therefore 2^{j-1} possible chains for j infections of m contacts. For $j, m = 4$, the possible chains are:

1. $\{1, 1, 1, 1\}$
2. $\{1, 0, 1, 1\} \to \{1, 1, 1, 1\}$
3. $\{1, 1, 0, 1\} \to \{1, 1, 1, 1\}$
4. $\{1, 1, 1, 0\} \to \{1, 1, 1, 1\}$
5. $\{1, 0, 0, 1\} \to \{1, 1, 0, 1\} \to \{1, 1, 1, 1\}$
6. $\{1, 0, 1, 0\} \to \{1, 1, 1, 0\} \to \{1, 1, 1, 1\}$
7. $\{1, 1, 0, 0\} \to \{1, 1, 1, 0\} \to \{1, 1, 1, 1\}$
8. $\{1, 0, 0, 0\} \to \{1, 1, 0, 0\} \to \{1, 1, 1, 0\} \to \{1, 1, 1, 1\}$

The associated probabilities for each chain (from the lecture slides), are given by:

$$\Pr(i_0 \to i_1 \to \cdots \to i_r) = \frac{s_0!}{i_1! i_2! \cdots i_r! s_r!} \prod_{t=0}^{r} p^{i_{t+1}} q^{s_{t+1}} \tag{1}$$

so the probability of each of the 8 chains for $m, j = 4$ is:

1. $\{1,1,1,1\}$
   $s_0 = 4,\ i_1 = 1,\ i_2 = 1,\ i_3 = 1,\ i_4 = 1,\ s_4 = 0$
   $\Pr(\{1,1,1,1\}) = \frac{4!}{1!1!1!1!0!} \cdot p^4 \cdot q^0 = 24 \cdot p^4$
2. $\{1,0,1,1\} \rightarrow \{1,1,1,1\}$
   $s_0 = 4,\ i_1 = 1,\ i_2 = 0,\ i_3 = 1,\ i_4 = 1,\ s_4 = 0$
   $\Pr(\{1,0,1,1\} \rightarrow \{1,1,1,1\}) = \frac{4!}{1!0!1!1!0!} \cdot p^3 \cdot q^1 = 24 \cdot p^3 q$
3. $\{1,1,0,1\} \rightarrow \{1,1,1,1\}$
   $s_0 = 4,\ i_1 = 1,\ i_2 = 1,\ i_3 = 0,\ i_4 = 1,\ s_4 = 0$
   $\Pr(\{1,1,0,1\} \rightarrow \{1,1,1,1\}) = \frac{4!}{1!1!0!1!0!} \cdot p^3 \cdot q^1 = 24 \cdot p^3 q$
4. $\{1,1,1,0\} \rightarrow \{1,1,1,1\}$
   $s_0 = 4,\ i_1 = 1,\ i_2 = 1,\ i_3 = 1,\ i_4 = 0,\ s_4 = 0$
   $\Pr(\{1,1,1,0\} \rightarrow \{1,1,1,1\}) = \frac{4!}{1!1!1!0!0!} \cdot p^3 \cdot q^1 = 24 \cdot p^3 q$
5. $\{1,0,0,1\} \rightarrow \{1,1,0,1\} \rightarrow \{1,1,1,1\}$
   $s_0 = 4,\ i_1 = 1,\ i_2 = 0,\ i_3 = 0,\ i_4 = 1,\ s_4 = 0$
   $\Pr(\{1,0,0,1\} \rightarrow \{1,1,0,1\} \rightarrow \{1,1,1,1\}) = \frac{4!}{1!0!0!1!0!} \cdot p^2 \cdot q^2 = 24 \cdot p^2 q^2$
6. $\{1,0,1,0\} \rightarrow \{1,1,1,0\} \rightarrow \{1,1,1,1\}$
   $s_0 = 4,\ i_1 = 1,\ i_2 = 0,\ i_3 = 1,\ i_4 = 0,\ s_4 = 0$
   $\Pr(\{1,0,1,0\} \rightarrow \{1,1,1,0\} \rightarrow \{1,1,1,1\}) = \frac{4!}{1!0!1!0!0!} \cdot p^2 \cdot q^2 = 24 \cdot p^2 q^2$
7. $\{1,1,0,0\} \rightarrow \{1,1,1,0\} \rightarrow \{1,1,1,1\}$
   $s_0 = 4,\ i_1 = 1,\ i_2 = 1,\ i_3 = 0,\ i_4 = 0,\ s_4 = 0$
   $\Pr(\{1,1,0,0\} \rightarrow \{1,1,1,0\} \rightarrow \{1,1,1,1\}) = \frac{4!}{1!1!0!0!0!} \cdot p^2 \cdot q^2 = 24 \cdot p^2 q^2$
8. $\{1,0,0,0\} \rightarrow \{1,1,0,0\} \rightarrow \{1,1,1,0\} \rightarrow \{1,1,1,1\}$
   $s_0 = 4,\ i_1 = 1,\ i_2 = 0,\ i_3 = 0,\ i_4 = 0,\ s_4 = 0$
   $\Pr(\{1,0,0,0\} \rightarrow \{1,1,0,0\} \rightarrow \{1,1,1,0\} \rightarrow \{1,1,1,1\}) = \frac{4!}{1!0!0!0!0!} \cdot p^1 \cdot q^3 = 24 \cdot pq^3$

Question 4

Greenwood presents his expected distribution of secondary measles cases from the chain binomial model with $m = 4$ in Table III. What values of p and q did he use to arrive at this expected distribution? Show how he derived these estimates.

Greenwood used what he termed "Newton's approximation", where he set $q^m = f_0/F$ where the f_0 is frequency of groups with no secondary infections, and F is the total number of groups. He performs the operation twice. To find the mean number of infected contacts for $m = 4$ he calculates is as the number of contacts infected times the number of households observed in which i infected out of m contacts were infected, divided by the number of households of that size. For $m = 4$, that's here:

```
h4 <- sum(infect_freq[, 5])
sum(infect * infect_freq[, 5]) / h4
```

```
## [1] 1.180328
```

Then he takes this value, sets it equal to the equation for the mean that he derives, which for $m = 4$ is

$$Mean = 4 - 4q^2 + 12q^3 - 40q^4 + 52q^5 - 24q^6 - 60q^7 + 132q^8 - 96q^9 + 24q^{10}$$

He then comes back to his equation for q, and for $m = 4$, $f0 = 60$ and $F = 132$ and so therefore,

$$\begin{aligned} q^4 &\approx 0.4545455, \text{ so,} && (2) \\ q &= \sqrt[4]{0.4545455} && (3) \\ & && (4) \end{aligned}$$

which we can calculate as:

```
0.4545455^(1 / 4)
```

```
## [1] 0.8210968
```

Which is our first approximation, $q \approx 0.821$, so $p \approx 0.179$

Question 5

Why might the values of p and q for households with m = 4 be expected to differ from those for households with m = 3 in this study? Give at least two reasons. (HINT: I don't want to know why the estimates might differ, but rather why the "true" values might differ. Think about what assumptions you are making and why they might be violated.)

The main assumption as to why the true values of different households may be different is the assumption of homogeneity of contacts, as well as the assumption of static contact timing / duration of exposure. One can imagine that in a household where two people are infected, if there is just one additional person ($m = 3$) then there's some assumption implicitly that the level of exposure of that third person is the same as if there were three people infected ($m = 4$) and the fourth person was exposed. There are any number of reasons why this could be considered an assumption that might not hold true. One can imagine that if one child in a household is infected (particularly of measles), then the household quarantine and/or care behaviours will likely be different in differently sized households. Additionally, the age structure likely plays a role that isn't accounted for between households. As the number of household contacts goes up, the likelihood increases that there is a larger age gap between contacts which may change the transmission probabilities since household contacts with larger age gaps are less likely to share bedrooms or be in close contact etc.

Question 6

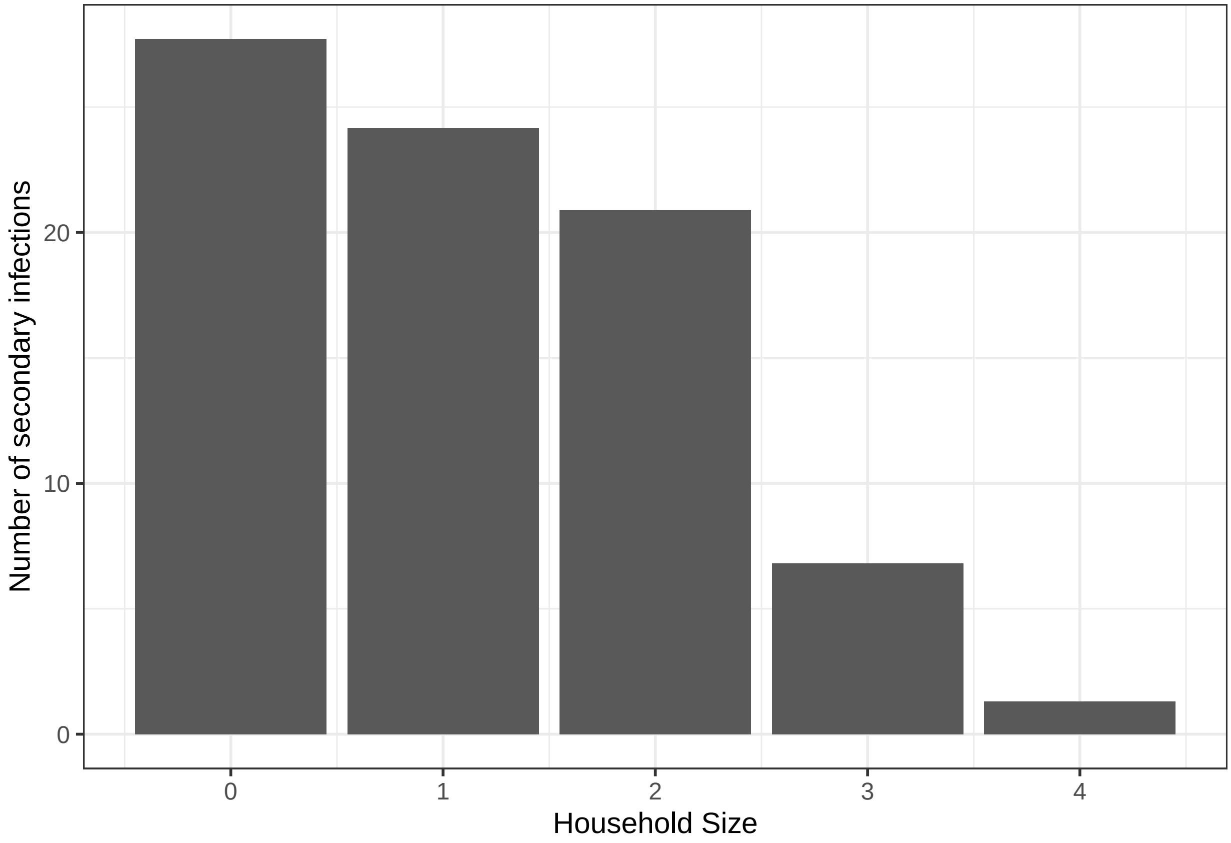
What would be the expected distribution of secondary cases for m = 4 under the Reed-Frost assumption given the values of p and q you calculated in question 4?

```
# Probability of each chain (under RF assumption)
q <- 0.821
p <- (1 - q)

p_rf_chain <- c(
    q^4,
    4 * p * q^3,
    6 * p^2 * q^2 + 12 * p^2 * q^3,
    4 * p^3 * q + 24 * p^3 * q^2,
    p^4 + 24 * p^4 * q
)
p_infectRF <- c(
    p_rf_chain[1],
    p_rf_chain[2],
    p_rf_chain[3],
    p_rf_chain[4],
    p_rf_chain[5]
)

e_inf_rf <- h4 * p_rf_chain

df_rf <- data.frame(
    sec_inf = h4 * p_rf_chain,
    hhs = c(0:4)
)
ggplot2::ggplot(df_rf) +
    ggplot2::geom_col(ggplot2::aes(x = hhs, y = sec_inf)) +
    ggplot2::labs(x = "Household Size", y = "Number of secondary infections") +
    ggplot2::theme_bw()
```



Question 7

EXTRA CREDIT: Which assumption (Greenwood vs Reed-Frost) provides a better fit to the observed distribution of secondary cases when $m = 4$? Show the statistical criteria you used to determine this.

We can do this with the simple mean-squared error:

```r
observed_data <- c(60, 29, 25, 11, 7)
expected_rf <- df_rf$sec_inf
expected_gw <- df_gw$sec_inf

mse_rf <- mean((observed_data - expected_rf)^2)
mse_gw <- mean((observed_data - expected_gw)^2)

print(paste0("Reed-frost MSE: ", mse_rf, ", Greenwood MSE: ", mse_gw))
```

```
## [1] "Reed-frost MSE: 226.549098182791, Greenwood MSE: 315.54206816"
```

So the better fit is Reed-frost.