**EMD 538**

**Lab 2**

**September 15, 2023**

In today’s lab, we are going to analyze the data from Greenwood, *J Hygiene* (1931) on the occurrence of measles in household contacts (children <10 years of age) of an index case during a 1926 outbreak in St. Pancras (London).

Let's consider the households with m=3 contacts.

1. What is the total number of households with m=3 contacts?

2. First, assume there is NO on-going transmission in households. Estimate p from the secondary attack rate (SAR) among ALL households (assuming that everyone who was infected had symptom onset within the maximum serial interval).

3. Calculate the expected frequency of secondary infections (for m=3) given NO on-going transmission.

Now let's estimate p and q for the chain binomial model.

4. Calculate the observed mean number of infected contacts for m=3.

5. Define the equation for the expected number of infected contacts, then solve.

Now let’s do this in R…

Input the data from Greenwood (1931).

Let's consider the households with m=3 contacts.

1. What is the total number of households (“c”) with m=3 contacts?

2. First, assume there is NO on-going transmission in households. Estimate p from the secondary attack rate (SAR) among ALL households (assuming that everyone who was infected had symptom onset within the maximum serial interval).

3. Calculate the expected frequency of secondary infections given NO on-going transmission.

HINT: Use dbinom

Now let's estimate p and q for the chain binomial model.

4. Calculate the observed mean number of infected contacts.

5. Use optimization to solve for the expected number of contacts (by minimizing the squared error)

6. Assign the value of “q” that is a REAL number between 0 and 1 and calculate “p”.

7. Calculate the probability of each possible chain, using Greenwood assumption.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Number of contacts infected** | **Possible chains** | **Probability**  **(Greenwood)** | **Probability**  **(Reed-Frost)** | **Probability**  **(in general)** |
| 0 | 1🡪0 | *q*3 | *q*3 | *q*13 |
| 1 | 1🡪1 | 3*pq*4 | 3*pq*4 | 3*p*1 *q*14 |
| 2 | 1🡪1🡪1 | 6*p2q*4 | 6*p2q*4 | 6*p*1*2 q*14 |
|  | 1🡪2 | 3*p*2*q*2 | 3*p*2*q*3 | 3*p*12 *q*1 *q*2 |
| 3 | 1🡪3 | *p*3 | *p*3 | *p*13 |
|  | 1🡪1🡪1🡪1 | 6*p*3*q*3 | 6*p*3*q*3 | 6*p*13 *q*13 |
|  | 1🡪2🡪1 | 3*p*3*q* | 3*p*2*q*(1-*q*2) | 3*p*12 *q*1 *p*2 |
|  | 1🡪1🡪2 | 3*p*3*q*2 | 3*p*3*q*2 | 3*p*13 *q*12 |

8. Calculate the probability of each possible number of secondary infections, using Greenwood assumption.

9. Calculate the expected distribution of secondary infections, using Greenwood assumption.

10. What if we follow the Reed-Frost assumption instead of the Greenwood assumption? Re-calculate the probability of each chain (using the same values of p and q).

Now let's relax our assumptions and estimate separate values for q1 and q2. We can make an initial "guess" at what the values of q1 and q2 are (see Becker (1989) p. 18).

Next, we can estimate q1 and q2 via maximum likelihood estimation.

Refresher: probability vs likelihood

Up to this point, we’ve more or less been operating under the assumption that we **know** *p, q,* and *m* (these make up our model parameters, which we’ll call ***θ***) and we want to **estimate** the probability of our observed data, which we’ll call ***X***. We can denote this Pr(***X***|***θ***), i.e. the **probability** of our data given the parameters. But if instead we **know** that we have observed some data ***X*** and we want to **estimate** the model parameters ***θ*** that gave rise to this data, in this case we refer to the **likelihood** of the parameters given the observed data: *L*(***θ***|***X***). To find the set of parameters that are *most* consistent with the data that we observed, we want to maximize the likelihood of the parameters given the data.

For example, if we flip a coin 10 times, and it comes up heads 8 out of 10 times, we can ask what is the **probability** of that outcome given *p*=0.5. Or we can ask what is the **likelihood** that it is a fair coin (i.e. *p*=0.5) given the outcome we observed. The answer is the same, but the question is different.

dbinom(8,10,0.5)

plot(dbinom(8,10,seq(0.01,0.99,by=0.01)))

example = function(p){-dbinom(8,10,p)}

optim(0.5,example)$par

logexample = function(p){-log(dbinom(8,10,p))}

optim(0.5,logexample)$par

Since there is no set way to calculate the likelihood for the chain binomial model in R, we can create a "function" file called 'chainbinolikl3' that calculates the negative log-likelihood for our model parameter(s) (***q***) and our observed data (***infect\_freq[,3]***).

11. Calculate the value of the negative log-likelihood at our initial “guesses” for q1 and q2.

12. Compare this to the negative log-likelihood for the “naïve” guess of q1=q2=0.5.

13. Now use optim() to minimize the value of the function in order to estimate q1 and q2:

14. What is the value of the negative log-likelihood at our best-fit estimate?