L07 COVID-19 Outbreak Investigations[1]

Michael Höhle[1]

[1]Department of Mathematics, Stockholm University, Sweden
🐦 m_hoehle

STA427 FS2021
Statistical Methods in Infectious Disease Epidemiology
Epidemiology, Biostatistics and Prevention Institute
University of Zurich, Switzerland

University of Zurich[UZH]

Outline

Outline

Outbreak Investigation

- In this course we have looked a lot at the use of statistical methods for modelling population level time series, i.e. surveillance data
- However, many infectious disease outbreaks are (or start as) as small localized clusters, which can be investigated by *field investigations*.
- Aims of such an outbreak investigation[2]:
  - Control or prevention (knowledge of agent, course of the outbreak, mode of transmission, source)
  - Research opportunity (mode of transmission, incubation period, clinical spectrum, ...)
  - Public, political, or legal concerns
  - Training

---

[2]Taken from https://www.cdc.gov/csels/dsepd/ss1978/lesson6/section1.html

Secondary attack rate

- The *secondary attack rate* is an alternative morbidity measure to, e.g. incidence, and is defined as[3]

$$\text{SAR} = \frac{\text{Number of new cases among contacts}}{\text{Total number of contacts}}$$

- Classical methods for calculating point estimate and CIs for a proportion can be used.
- By calculating the SAR in two groups one can even compute relative risks by classical 2x2 methods.

---

[3] https://www.cdc.gov/csels/dsepd/ss1978/lesson3/section2.html

Outline

1. Outbreak Investigations

2. Example: COVID-19 Carnival cluster, Germany, Feb-Mar 2020

3. Transmission Graphs

4. Discussion

COVID-19 Carnival cluster (1)

- Early COVID-19 outbreak investigation by Bender et al. (2021) in order to determien SAR as well as incubation period, serial interval and generation time.
- Important insight: There seems to be little transmission from completely asymptomatic cases
- SAR for household contacts of lab confirmed cases:

Clinical symptoms source	No. contacts infected[4]	Total no. contacts	SAR
Asymptomatic	0	4	0%
Symptomatic[5]	4	28	14.3%
Total	4	32	12.5%

---

[4] either tested positive or experienced respiratory symptoms

[5] phase not specified or both

COVID-19 Carnival cluster (2)

- Getting an OR estimate with CI for the SAR of asymptomatic vs. symptomatic is difficult, due to the zero in the 2×2 table

```
## infected n   cs_source
## 1      0  4 asymptomatic
## 2      4 28  symptomatic

m_glm <- glm( cbind(infected, n-infected) ~ 1 + cs_source, data=carnival, family=binomial )
confint(m_glm)
## Waiting for profiling to be done...
##                         2.5 %   97.5 %
## (Intercept)                NA 4097.869
## cs_sourcesymptomatic -3292.433      NA
```

- Problem with zero can be addressed by *exact logistic regression* as implemented in the elrm package (Zamar et al. 2007):

```
#devtools::install_github(repo="https://github.com/cran/elrm.git")
m <- elrm::elrm( infected/n ~ cs_source, interest=~cs_source, dataset=carnival, r=2, iter=1.5e4, burnIn=5e3)

c(hat=as.numeric(exp(m$coeffs)), exp(m$coeffs.ci)) %>% unlist
##        hat      lower      upper
## 0.78298817 0.07570849        Inf
```

- OR estimate and CI is close to the RR estimate stated in Bender et al. (2021)

COVID-19 Carnival cluster (3)

- Same for other contacts of lab confirmed cases

| Clinical symptoms source	No. contacts infected[6]	Total no. contacts	SAR
Asymptomatic cases	0	22	0%
Symptomatic[7]	3	25	12.0%
Symptomatic, presymptomatic phase only	15	72	20.8%
Symptomatic, symptomatic phase only	2	29	6.9%
Total	20	148	13.5%

---

[6] either tested positive or experienced respiratory symptoms

[7] phase not specified or both

Outline

1. **Outbreak Investigations**

2. **Example: COVID-19 Carnival cluster, Germany, Feb-Mar 2020**

3. **Transmission Graphs**
   - **Interval Censoring**
   - **Multiple trees**

4. **Discussion**

Outbreak Investigations
○○○

Carnival cluster
○○○○

Transmission Graphs
○●○○○○○○○○○

Discussion
○○

References

Transmission Graph(s)
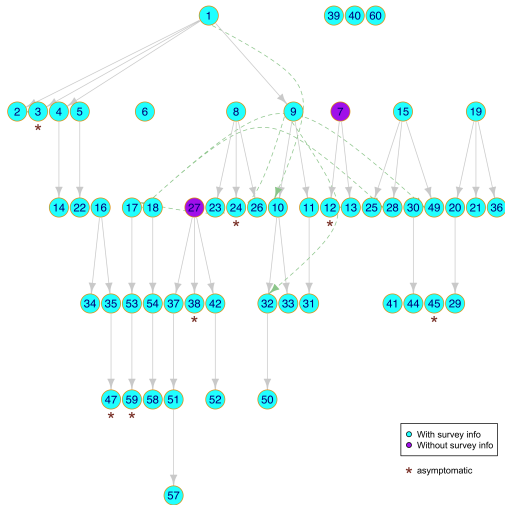
Graph Notation (1)

- We describe the outbreak for the n individuals by the transmission tree $T = (V, E)$. By $v \in V$ with $|V| = n$ denote the nodes of the directed graph and E denotes the set of directed edges.
- For a node v let $\mathrm{pa}(v)$ and $\mathrm{ch}(v)$ denote the set of parents and children, respectively. Since T is a tree $|\mathrm{pa}(v)| = 1$ for any v except the root, which is the node which has $\mathrm{pa}(v) = \emptyset$.
- Assuming the outbreak is observed until its end, the average number of secondary cases a primary cases generates is

$$\frac{1}{|V|} \sum_{v \in V} |\mathrm{ch}(v)|.$$

Graph Notation (2)

- Let $\mathrm{dso}(v)$ denote the day of symptom onset of $v \in V$ and let p_v denote v's parent in T. The empirical distribution of the serial interval time is formed by the values $\mathrm{dso}(v) - \mathrm{dso}(p_v)$, for all v nodes in the set of nodes with known source, i.e.

$$V_p = \{v \in V : |\mathrm{pa}(v)| = 1\}.$$

- Let $e(v)$ be the time of exposure of $v \in V$. The empirical *generation time distribution* is formed by the values $t(v) - e(p_v)$, for all nodes $v \in V_p$.

- The empirical *incubation period distribution* is given by the values $\mathrm{dso}(v) - e(v)$ for all nodes $v \in V$.

Interval Censoring (1)

- As noted by Reich et al. (2009) the exact timing of exposition or symptom onset are not always given $\rightarrow$ doubly interval censored data
- If instead the time of exposure and the onset of symptoms are only known to fall within a finite interval, then a typical observation for an infector-infectee pair consist of

$$X(v) = (e(v)_L, e(v)_R, \mathrm{dso}(v)_L, \mathrm{dso}(v)_R).$$

- Illustration:
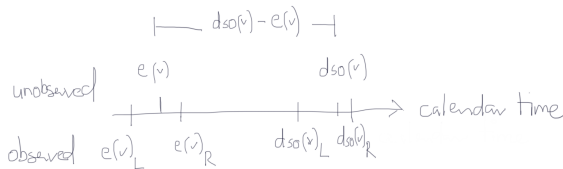
Interval Censoring (2)

- We follow the approach by Reich et al. (2009).
- Let the incubation period T be a non-negative continuous random variable with PDF $f_\theta(t)$ and let $h_\lambda(e)$ to be the PDF of the infecting exposure time E, dependent on a parameter λ, and let $g(s)$ to be the PDF of the DSO (in calendar time). Assume E to be independent of the incubation period T.
- We have that

$$g(s|e) = f_\theta(s-e|e) = f_\theta(s-e).$$

- The joint PDF of E and the dso is given by

$$p(e,s) = p(e)p(s|e) = h_\lambda(e)g(s|e) = h_\lambda(e)f_\theta(s-e).$$

Interval Censoring (3)

- The likelihood for a single doubly interval-censored observation is therefore

$$L(\theta, \lambda; X) = \int_{E_L}^{E_R} \int_{\mathrm{dso}_L}^{\mathrm{dso}_R} h_\lambda(e) f_\theta(s-e) ds \, de.$$

- Typically, a parametric accelerated failure time model is used to model the incubation period, e.g., a Log-Normal distribution.
- An implementation of this approach exists in the R package `coarseDataTools` (Reich et al. 2019)
- Can even extend the modelling with covariates such as age or sex (Werber et al. 2013)

Multiple trees (1)

- As seen from the transmission tree, it is not always 100% clear who infected who. Instead, up to 3 potential sources could be specified for each case.
- Let $|\operatorname{pa}(v)| \geq 1$, $v \in V_p$, denote the set of possible sources in the graph representing transmissions.
- A transmission tree T can again be obtained by selecting one distinct parent node v_p for each node v with $|\operatorname{pa}(v)| > 1$.
- Let $\mathcal{T}$ denote the set of possible transmission trees, i.e.

$$|\mathcal{T}| = \prod_{v \in V_p} |\operatorname{pa}(v)|.$$

Multiple trees (2)

- Simple way to average estimation of incubation period, serial interval and generation time over all possible trees:
  - Non-parametric case: for each $T \in \mathcal{T}$ compute the empirical distribution based on the n cases and compute quantiles, mean, etc. based on the set of $n \times |\mathcal{T}|$ values
  - Parametric case: for each $T \in \mathcal{T}$ draw k samples from the respective estimated parametric distribution and compute quantiles, mean, etc. based on the set of $k \times |\mathcal{T}|$ values
- Note: The above averaging approach weighs each $T \in \mathcal{T}$ equally. In principle some trees are more plausible than others $\rightarrow$ Likelihood or Bayesian framework for data augmentation

Results

- Results from Bender et al. (2021) for serial interval (non-parametric), generation time (parametric) and incubation period (parametric)
- A pre-processing step was applied to symptomatic cases filtering out possible sources if they did not meet within 2 days before DSO or 10 days after DSO.
- Results averaged over the 144 possible transmission trees:

Time	1%	5%	25%	50%	75%	95%	99%	Mean
Serial interval (d)	-2.0	-1.0	1.0	3.0	6.0	15.0	22.0	4.5
Generation time (d)	0.1	0.3	1.7	3.6	6.6	13.1	21.6	4.9
Incubation period (d)	0.3	0.8	2.5	4.3	6.5	10.6	14.3	4.8

(Quantiles)

- Note: The serial interval is rather short an can actually be negative. Shows why COVID-19 is hard to control.

Outline

1 Outbreak Investigations

2 Example: COVID-19 Carnival cluster, Germany, Feb-Mar 2020

3 Transmission Graphs

4 Discussion

Discussion

- Outbreak investigations are an important tool in the epidemiological toolbox, requires shoe-and-leather epidemiology
- Proper statistical methods as well as software toolboxes, e.g. in the form of R-packages, are needed more than ever! (Höhle 2017)
- For an overview of COVID-19 outbreak investigations in Germany see, e.g., Alpers et al. (2021) (in German)
- Other important COVID-19 investigations are, e.g., Russell et al. (2020), Yamagishi et al. (2020) and Murphy et al. (2020).

Literature I

📄 Alpers, Katharina, et al. 2021. "Untersuchung von
   SARS-CoV-2-Ausbrüchen in Deutschland durch Feldteams des Robert
   Koch-Instituts, Februar–Oktober 2020".
   Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz.

📄 Bender, Jennifer, Michael Brandl, Michael Höhle, Udo Buchholz, and
   Nadine Zeitlmann. 2021. "Analysis of Asymptomatic and
   Presymptomatic Transmission in SARS-CoV-2 Outbreak, Germany,
   2020". 27 (4).

📄 Höhle, M. 2017. "A Statistician's Perspective on Digital Epidemiology".
   Life Sciences, Society and Policy 13 (17).
   doi:10.1186/s40504-017-0063-9.

📄 Murphy, Nicola, et al. 2020. "A large national outbreak of COVID-19
   linked to air travel, Ireland, summer 2020". Eurosurveillance 25 (42).

Literature II

📄 Reich, Nicholas G, Justin Lessler, and Andrew S Azman. 2019.
coarseDataTools: A collection of functions to help with analysis of coarsely ob
R package version 0.6-5.
https://cran.r-project.org/package=coarseDataTools.

📄 Reich, Nicholas G, Justin Lessler, Derek A T Cummings, and
Ron Brookmeyer. 2009. "Estimating incubation period distributions
with coarse data." Stat Med 28, no. 22 (): 2769–2784.

📄 Russell, Timothy W, et al. 2020. "Estimating the infection and case
fatality ratio for coronavirus disease (COVID-19) using age-adjusted
data from the outbreak on the Diamond Princess cruise ship, February
2020". Eurosurveillance 25 (12).

📄 Werber, D., et al. 2013. "Associations of Age and Sex on Clinical
Outcome and Incubation Period of Shiga toxin-producing Escherichia
coli O104:H4 Infections, 2011". American Journal of Epidemiology 178
(6): 984–992.

Literature III

📄 Yamagishi, Takuya, Hajime Kamiya, Kensaku Kakimoto, Motoi Suzuki, and Takaji Wakita. 2020. "Descriptive study of COVID-19 outbreak among passengers and crew on Diamond Princess cruise ship, Yokohama Port, Japan, 20 January to 9 February 2020". <u>Eurosurveillance</u> 25 (23).

📄 Zamar, David, Brad McNeney, and Jinko Graham. 2007. "elrm: Software Implementing Exact-like Inference for Logistic Regression Models". <u>Journal of Statistical Software</u> 21 (3). `http://www.jstatsoft.org/`.