Predict extremes: influenza epidemics in France

joint work with Holger Rootzén (Chalmers University of Technology, Sweden)

Maud Thomas

LPSM, Sorbonne Université

Groupe de travail modélisation Covid-19

Sorbonne Université - April 14, 2020

Influenza

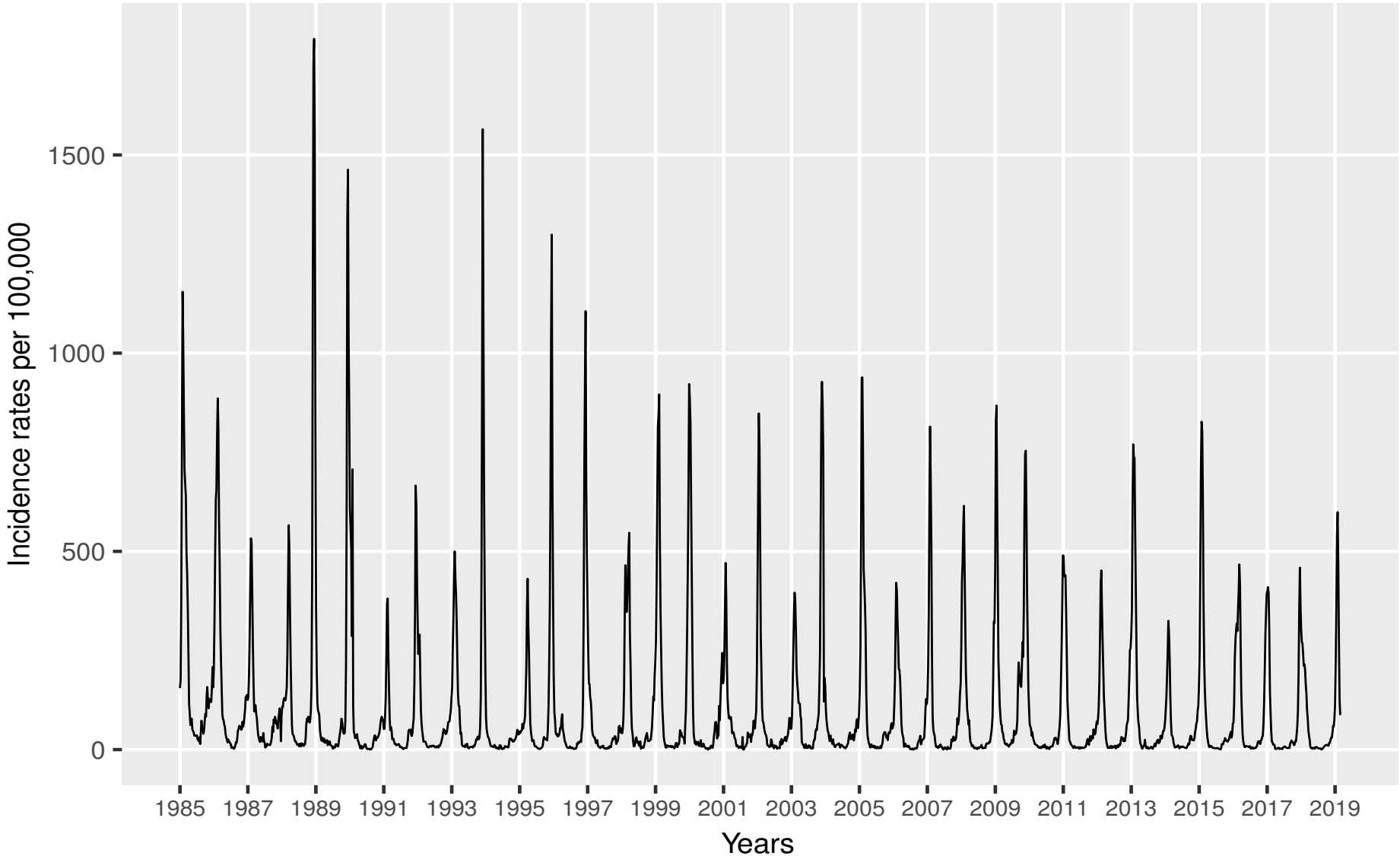
- Highly contiguous disease
  ↪ High morbidity
  ↪ High mortality
    - 250,000–500,000 deaths/year worldwide
    - 0–10,000 deaths/year in France
- High rate of mutation of influenza viruses ⇒ annual epidemics
- Concern for resource planning in public health
- DO NOT know the exact number of infected individuals or number of deaths

- ILI = Influenza-like illness
  ↪ fever above 39°C, muscle aches and respiratory symptoms
  ↪ good proxy for influenza incidence
- Incidence = number of new cases
- Sentinel network: weekly ILI incidence rates per 100,000 in France between January 1985 and February 2019
  ↪ 35 epidemics

Goal

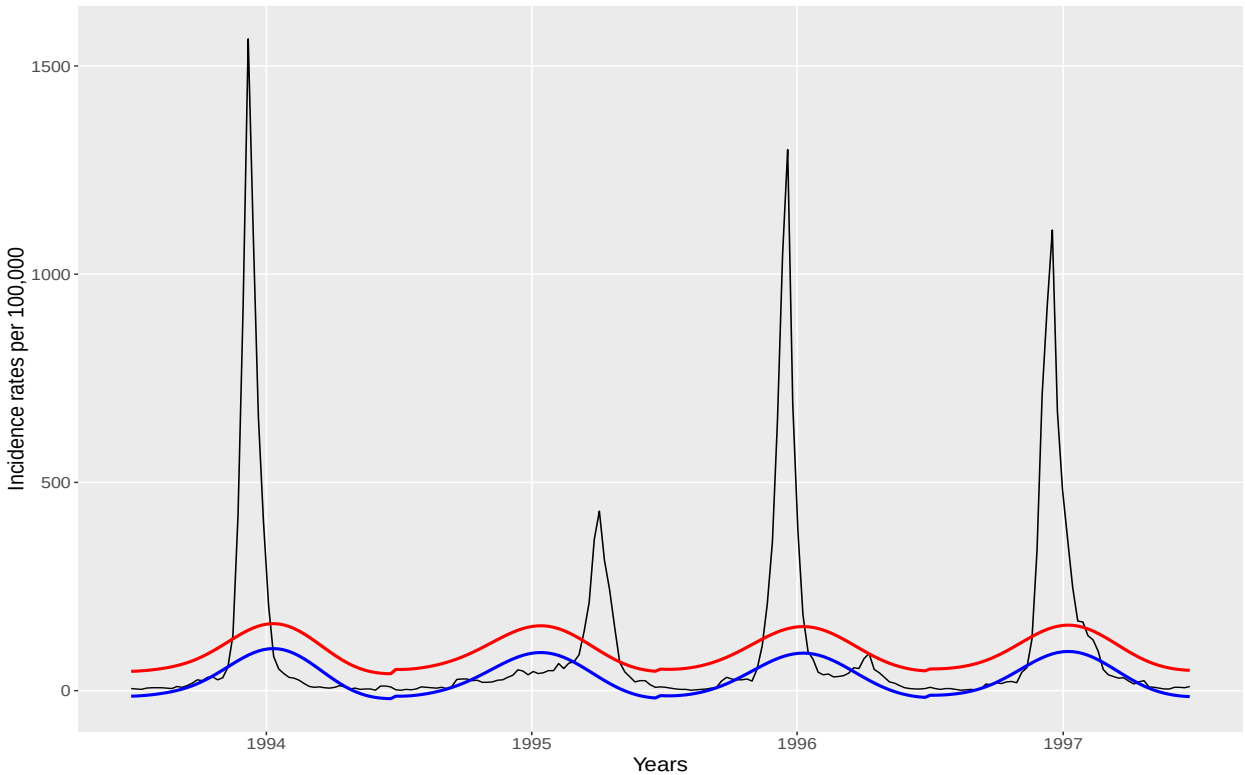
Predict the morbidity burden of exceptional or extreme influenza epidemics

ILI incidence rates in France between 1985 and 2019



Detection of epidemics

- Definition of an epidemic week: depends on the detection method
- **Serfling (1963)**: method used by the Sentinelles network
  ↪ Cyclic regression
  ↪ Epidemic week: if the incidence exceeds the 90%-upper bound of the prediction interval



Extreme flu epidemic?

- Extreme event = a value which exceeds some high threshold
- BUT often not ONE extreme event that is important, but consecutive ones

  Examples

  - Heat waves: sequence of very high minimum nightly temperatures
  - Extreme rainfalls: one day of very extreme rainfall, OR smaller, but still extreme rain amounts during several days
- Extreme flu epidemics:
  - one week with very extreme incidence rate,
  - consecutive weeks with quite high incidence rates

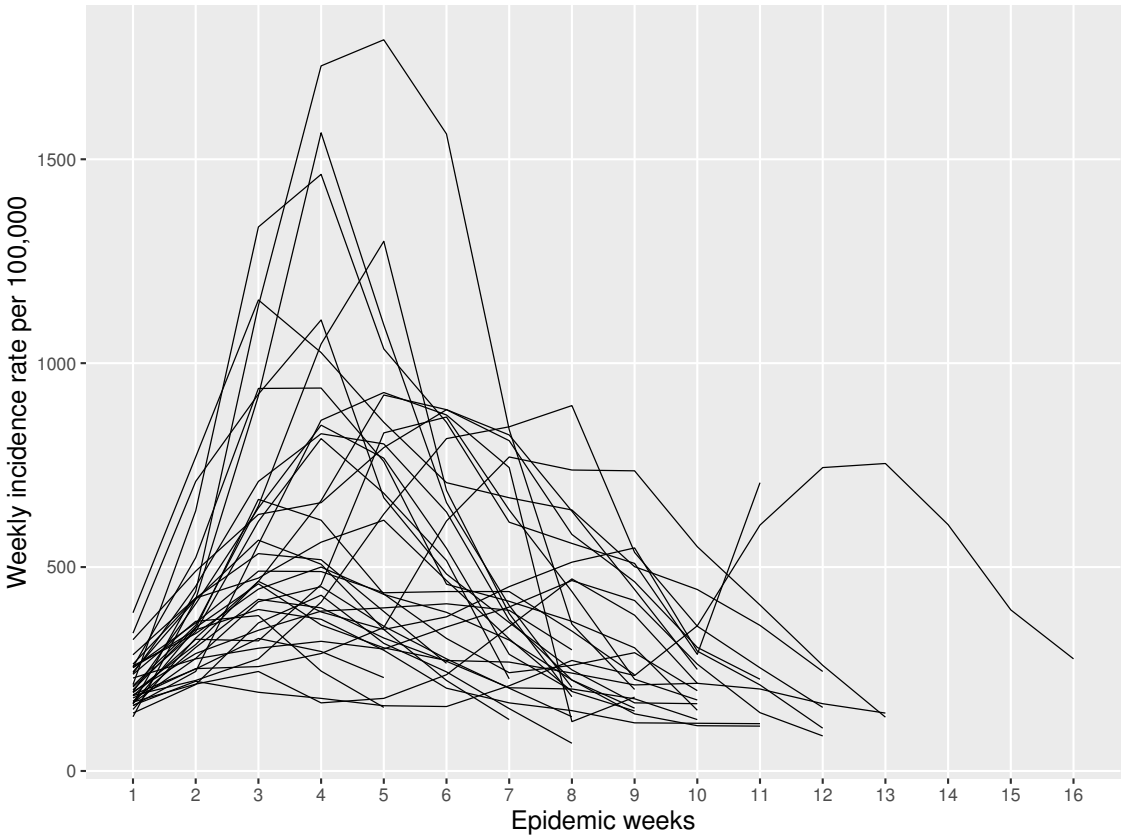
Key problems

(P1) estimation of risks of very extreme ILI rates in the next few years

(P2) real-time prediction of risks of high ILI rates in extreme epidemics

(P3) detection of anomalous and unusual, potentially very dangerous, extreme epidemics.

Synchronisation of epidemics

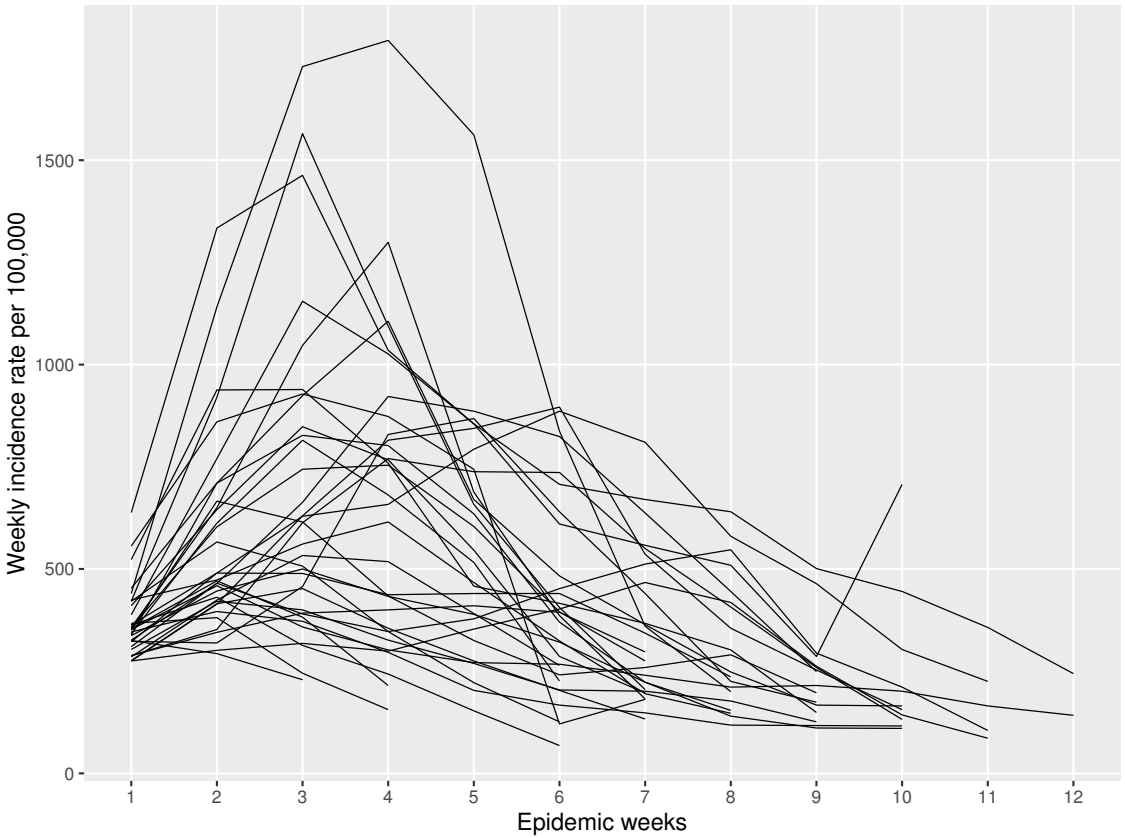


- Shortest: 5 weeks (1991)
- Longest: 16 weeks (2010)
- Highest peak: 1793 (1989)
- Smallest peak: 325 (2014)

$\Rightarrow$ Need a definition of the epidemic which makes epidemics synchronised

Only consider epidemic weeks when the incidence rate has reached a certain threshold

Synchronisation of epidemics



- t_{epid} = 272 per 100,000
- Exceedance two weeks in a row

- Y_1 = incidence rate of the first week above t_{epid}
- Y_2 = incidence rate of the second week above t_{epid}
- Y_3 = incidence rate of the week after Y_2
- ...
- Until the end of the Serfling epidemic period
- S = size of the epidemic
  $S = Y_1 + Y_2 + ...$

Goal

Predict of the severity of the epidemic given Y_1 and Y_2.

Key problems

- **Problem (P1)**
  For a (very) small probability α, estimation of $x_\alpha(m)$ such that Y_3 (or S,) will exceed $x_\alpha(m)$ with probability α in the next m years
- **Problem (P2)**
  For a high threshold v_3, estimation of

$$\mathbb{P}\left[Y_3 \geqslant v_3 \mid Y_1 = y_1, Y_2 = y_2\right]$$

  and

$$\mathbb{P}\left[S \geqslant v_3 \mid Y_1 = y_1, Y_2 = y_2\right]$$

- **Problem (P3)** based on the likelihood of the fitted model for (P2)

Key problems

- **Problem (P1)**
  For a (very) small probability α, estimation of $x_\alpha(m)$ such that Y_3 (or S,) will exceed $x_\alpha(m)$ with probability α in the next m years
- **Problem (P2)**
  For a high threshold v_3, estimation of

$$\mathbb{P}\left[Y_3 \geqslant v_3 \mid Y_1 = y_1, Y_2 = y_2\right]$$

  and

$$\mathbb{P}\left[S \geqslant v_3 \mid Y_1 = y_1, Y_2 = y_2\right]$$

- **Problem (P3)** based on the likelihood of the fitted model for (P2)

- v_3 can be larger than the observed maximum and p very small

$\Rightarrow$ Need inference outside of the support of the sample

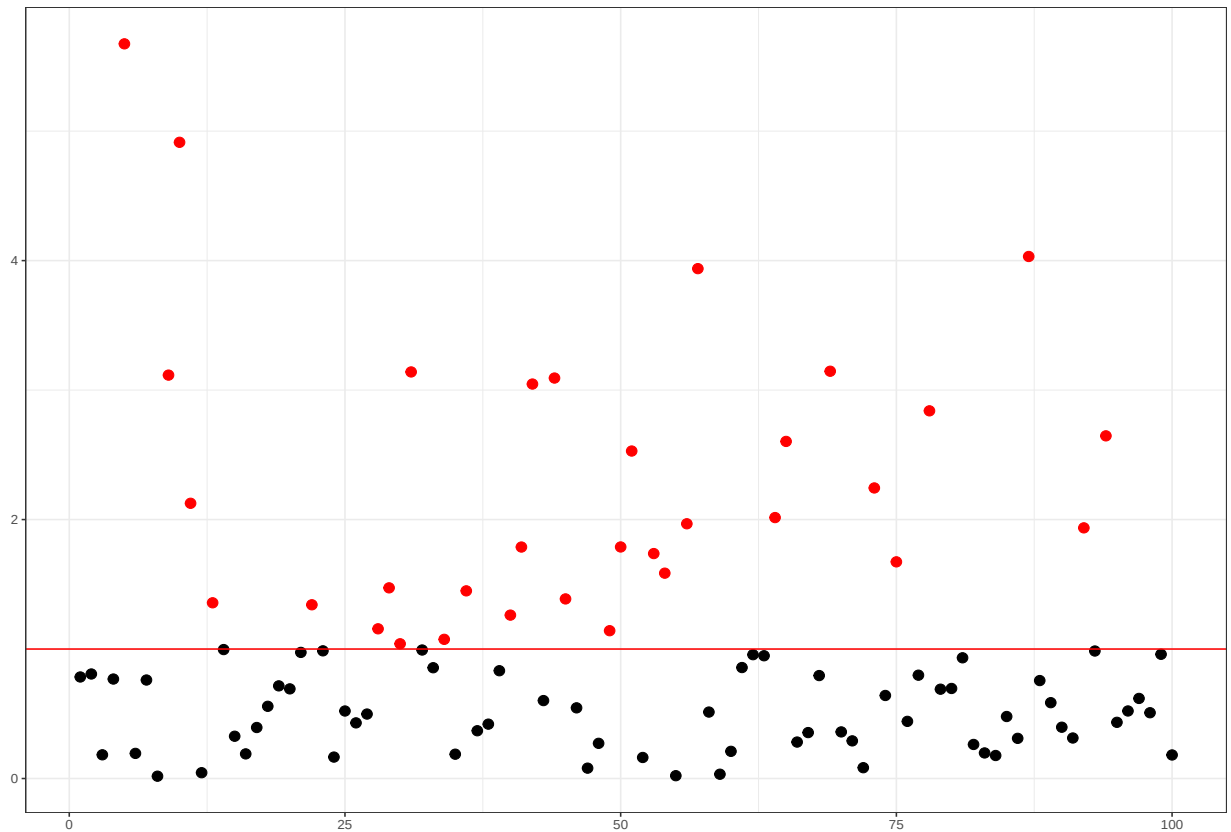
$\Rightarrow$ Extreme value theory

Extreme value theory

Univariate case

- $Y_1, Y_2, \ldots$ series of random variables
- Fix a (high) threshold u
- Extreme event = Y_i exceeds u

  → Given that $Y_i > u$, define the excess $X_i = Y_i - u$



Extreme value theory

Univariate case

- $Y_1, Y_2, \ldots$ series of random variables
- Fix a (high) threshold u
- Extreme event = Y_i exceeds u

  → Given that $Y_i > u$, define the excess $X_i = Y_i - u$

Balkema and de Haan (1974)

If there exist $(a_u) > 0$, (b_u) and a non-degenerated distribution function H such that, under conve

$$\mathbb{P}[Y_i - u \geqslant a_u x + b_u \mid Y_i > u] \xrightarrow[u\to\infty]{d} 1 - H(x)\,,$$

then H is necessarily of the form

$$H_{\sigma,\gamma}(x) = \begin{cases} 1 - \left(1 + \frac{\gamma}{\sigma} x\right)^{-1/\gamma} & \text{if } \gamma \neq 0 \\ 1 - \exp\left(-\frac{x}{\sigma}\right) & \text{if } \gamma = 0 \end{cases}$$

- Possible limits of excesses = Parametric family of distributions

  ↪ Generalized Pareto Distributions

(P1) Estimation of the m-year risk

- $\sigma > 0$ is a scale parameter
- $\gamma \in \mathbb{R}$ is a shape parameter
  → estimated by $\widehat{\sigma}$ and $\widehat{\gamma}$ (e.g. via MLE)
  → $H_{\sigma,\gamma}$ estimated by $\widehat{H} = H_{\widehat{\sigma},\widehat{\gamma}}$

- For $x > u$, $F(x) = \mathbb{P}[Y_i \geqslant x]$ estimated by

$$\widehat{F}(x) = 1 - \widehat{p}_u \left(1 - \widehat{H}(x-u)\right)$$

  where $\widehat{p}_u$ = estimation of the probability to have an extreme episode
- Cdf of the largest ILI rate during m years exceed x estimated by

$$\widehat{F}(x)^m = \left(1 - \widehat{p}_u \left(1 - \widehat{H}(x-u)\right)\right)^m$$

- The m-year risk $x_\alpha(m) = (1-\alpha)$ quantile of $\widehat{F}(x)^m$ given by inverting

$$\left(1 - \widehat{p}_u \left(1 - \widehat{H}(x-u)\right)\right)^m = 1 - \alpha$$

Application to the ILI data: marginal fitting

- Keep 1985-2018 data and put 2019 aside as a test case
- Apply PoT method to each marginal

$$
\begin{aligned}
Y_1^{1991}, \ldots, Y_1^{2018} &\rightarrow \text{First week of each epidemic} \\
Y_2^{1991}, \ldots, Y_2^{2018} &\rightarrow \text{Second week of each epidemic} \\
Y_3^{1991}, \ldots, Y_3^{2018} &\rightarrow \text{Third week of each epidemic} \\
\text{OR} & \\
S^{1991}, \ldots, S^{2018} &\rightarrow \text{Size of each epidemic}
\end{aligned}
$$

- Choose a GP threshold for each margin: $u = 0.9$-quantile (0.6-quantile for Size)

	Week 1	Week 2	Week 3	Size
GP threshold	339	339	339	4,145

(P1) Application to the ILI data: marginal fitting

- Fit an univariate GPD to excesses of each margin
- Likelihood ratio tests of the hypotheses that all γ are equal to 0

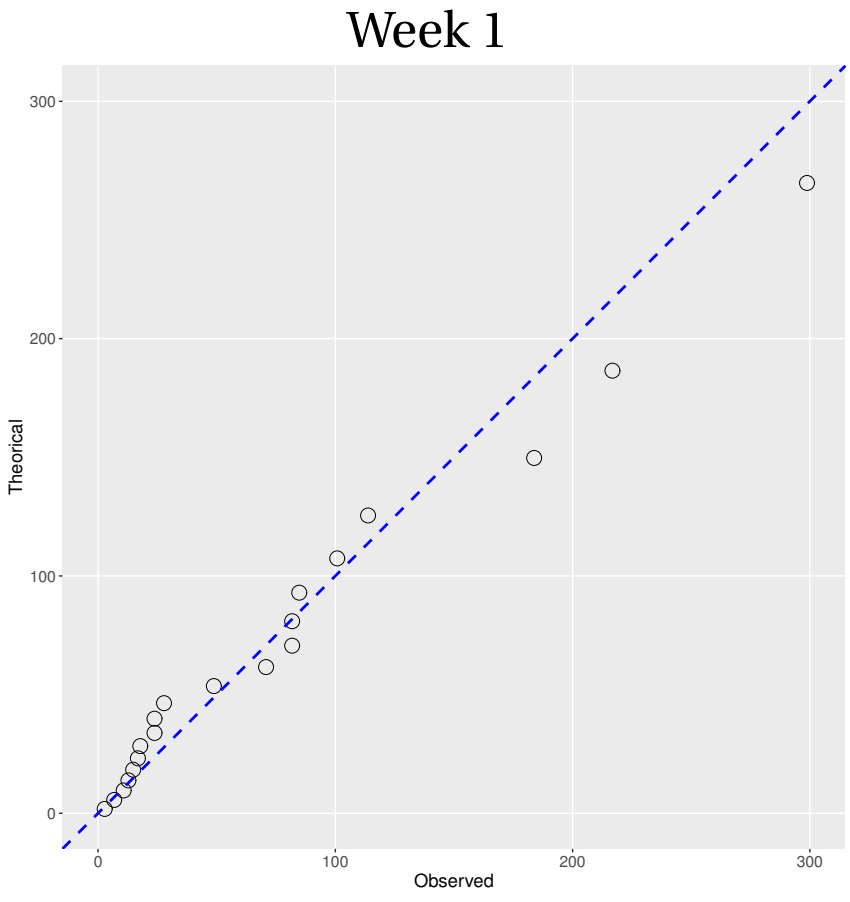
	Week 1	Week 2	Week 3	Size
p-value	0.66	0.57	0.64	0.98

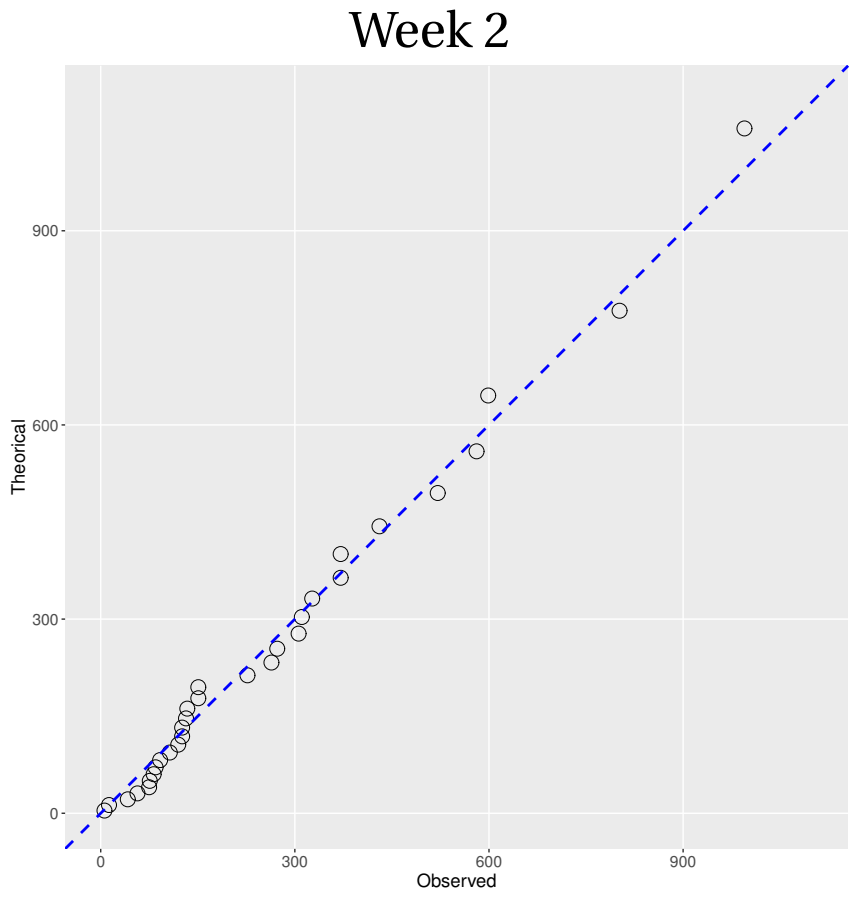
→ Assume that $\gamma = 0$

- Fit an exponential distribution

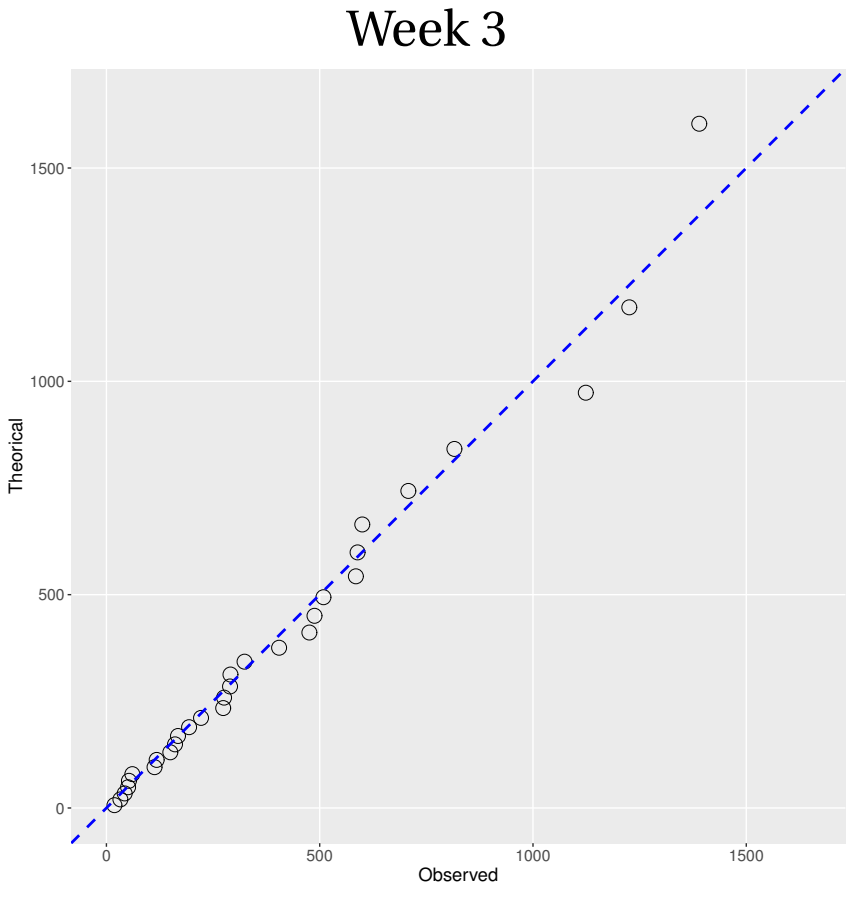
	Week 1	Week 2	Week 3	Size of epidemic
Estimate	72	256	392	1,428
95% CI	[40 ; 103]	[166 ; 347]	[251 ; 532]	[680 ; 2,175]

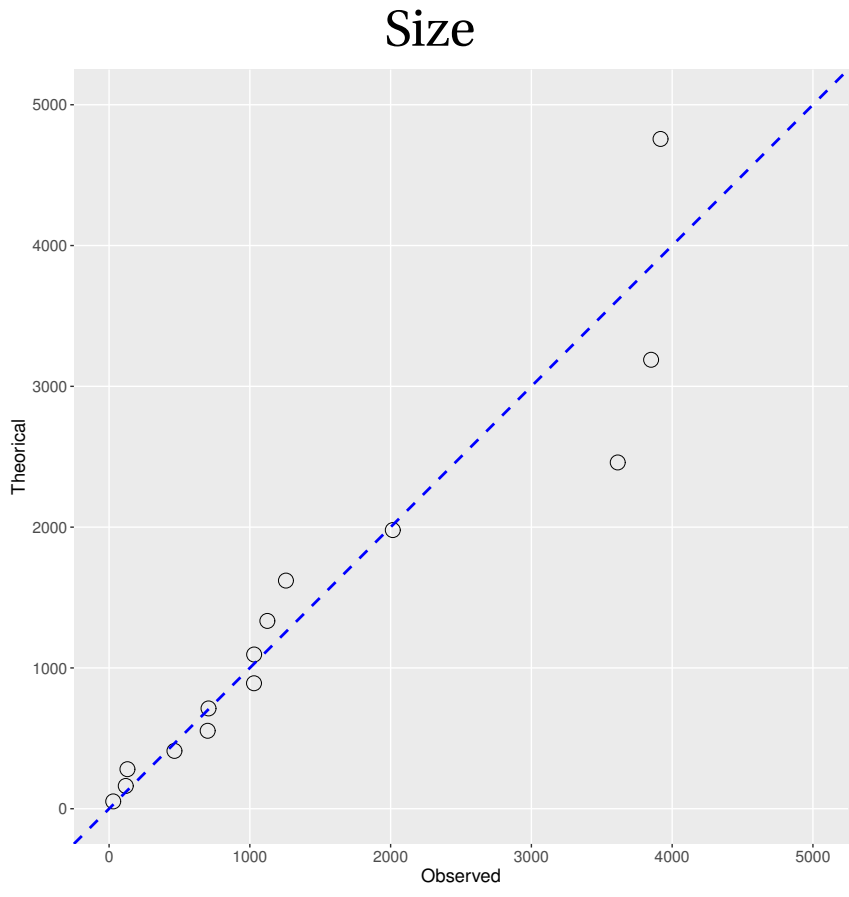
(P1) Marginal fitting exponential QQplots





(P1) Marginal fitting exponential QQplots





(P1) Application to the ILI data

Risks estimation

- Recall we need to inverse

$$\left(1-\widehat{p}_u\left(1-\widehat{H}(x-u)\right)\right)^m = 1-\alpha$$

- Here, $1-\widehat{H}(x) = \exp(-x/\widehat{\sigma})$

$$x_\alpha(m) = u + \widehat{\sigma}\{\log \widehat{p}_u - \log(1-(1-\alpha)^{1/n})\},$$

- And, $\widehat{p}_u$ is estimated by the empirical frequency, i.e., $30/34 \approx 0.88$ and $24/34 \approx 0.41$ for the Size

m	one year	one year	10 years	10 years
α	10%	1%	10%	1%
Week 3	1,192	2,094	2,076	2,994
Size of epidemic	6,165	9,452	9,385	12,733

(P1) Application to the ILI data

Risks estimation

- Recall we need to inverse

$$\left(1-\widehat{p}_u\left(1-\widehat{H}(x-u)\right)\right)^m = 1-\alpha$$

- Here, $1-\widehat{H}(x) = \exp(-x/\widehat{\sigma})$

$$x_\alpha(m) = u + \widehat{\sigma}\{\log \widehat{p}_u - \log(1-(1-\alpha)^{1/n})\},$$

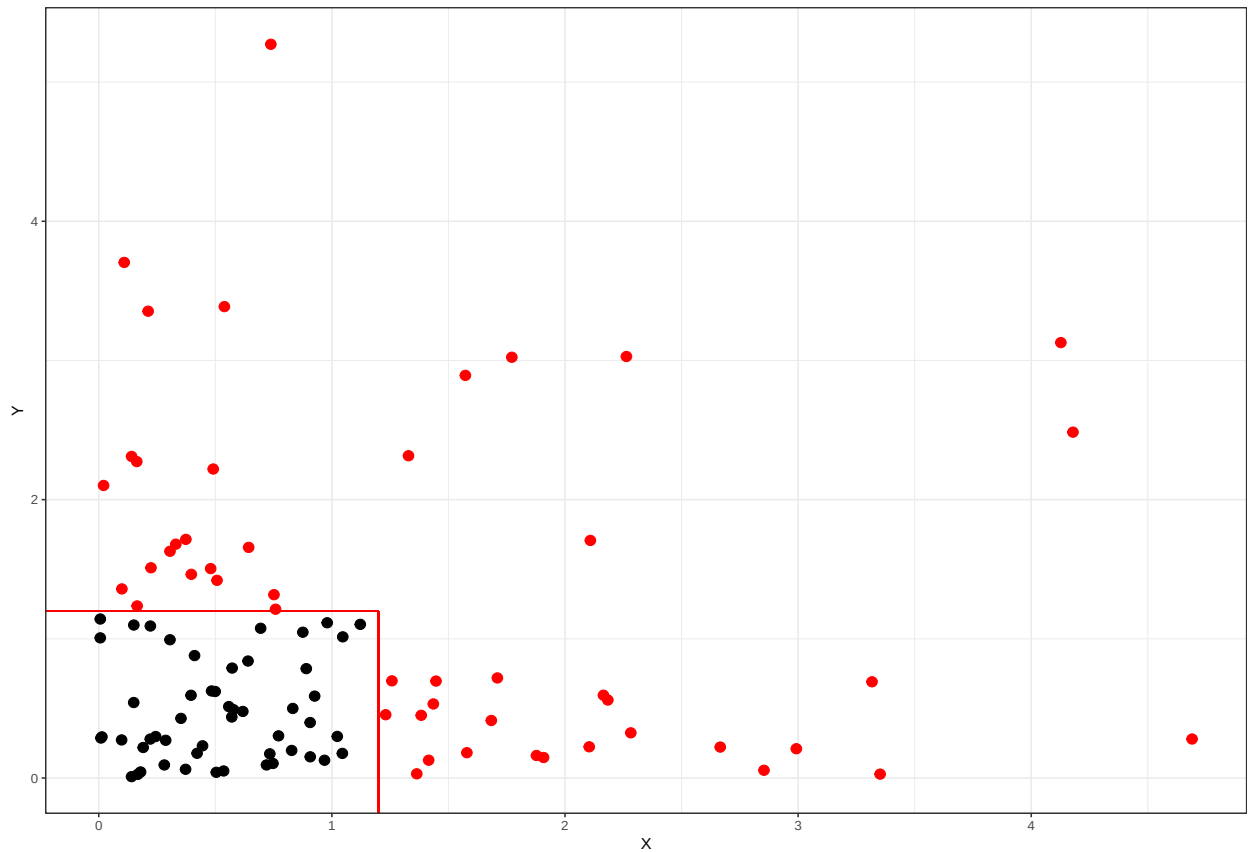
- And, $\widehat{p}_u$ is estimated by the empirical frequency, i.e., $30/34 \approx 0.88$ and $24/34 \approx 0.41$ for the Size

m / α	one year 10%	one year 1%	10 years 10%	10 years 1%
Week 3	1,192	2,094	2,076	2,994
Size of epidemic	6,165	9,452	9,385	12,733

Problem also addressed in Thomas et al. (2016)

Multivariate Generalized Pareto Distributions

- $\mathbf{Y} = (Y_1, Y_2, Y_3)$ observations
- Choose (high) thresholds $\mathbf{u} = (u_1, u_2, u_3)$
- Extreme event = AT LEAST one of the Y_j exceeds its threshold u_j



Multivariate Generalized Pareto Distributions

- $\mathbf{Y} = (Y_1, Y_2, Y_3)$ observations
- Choose (high) thresholds $\mathbf{u} = (u_1, u_2, u_3)$
- Extreme event = AT LEAST one of the Y_j exceeds its threshold u_j

NO parametric family of limits distributions

Multivariate Generalized Pareto Distributions

Representation of MGP vectors

- $\boldsymbol{U} = (U_1, U_2, U_3)$ such that $\mathbb{E}[e^{\max \boldsymbol{U}}] < \infty$
- $f_{\boldsymbol{U}}$ = density of $\boldsymbol{U}$

U-representation of MGPD vector (Kiriliouk et al. (2018))

Density of a MGPD with standard exponential margins generated by $\boldsymbol{U}$ is given by

$$h_{\boldsymbol{U}}(\boldsymbol{x}) = \frac{1_{\{\boldsymbol{x} \not\leq \boldsymbol{0}\}}}{\mathbb{E}[e^{\max \boldsymbol{U}}]} \int_0^\infty f_{\boldsymbol{U}}(\boldsymbol{x} + \log t) \mathrm{d}t,$$

$\Rightarrow$ Why is this interesting?

- Different distributions for $\boldsymbol{U}$
  $\hookrightarrow$ more different models
- Model defined via $\boldsymbol{U}$ have nice properties moving across dimension
  $\hookrightarrow$ for conditional prediction

(P2) Application to the ILI data: multivariate fitting

- **Excess:** $Y_j - u_j$ pour $j = 1, 2, 3$
- Standardized the data

$$X_j = \frac{Y_j - u_j}{\sigma_j} \sim \text{Exp}(1)$$

- Consider the exceedances

$$\begin{aligned} \boldsymbol{X}^{1985} &= \left(X_1^{1985}, X_2^{1985}, X_3^{1985}\right) \\ \boldsymbol{X}^{1986} &= \left(X_1^{1986}, X_2^{1986}, X_3^{1986}\right) \\ &\vdots \\ \boldsymbol{X}^{2018} &= \left(X_1^{2018}, X_2^{2018}, X_3^{2018}\right) \end{aligned}$$

$\Rightarrow$ Only consider $\mathbf{X}^i$ that are extreme, i.e., such that at least one of its component is above its threshold

$\rightarrow$ 32 for Week 3 and Size

(P2) Application to the ILI data: multivariate fitting

Parameter estimation

- **Model selection** → U_j, $j = 1, 2, 3$ independent $\sim$ Gumbel(α_j, β_j)
- For identifiability, β_1 set equal to 0
- Estimation of parameters for Week 3

Parameter	α_1	α_2	α_3	β_2	β_3
Estimate	2.22	10.37	3.21	0.84	0.59

- Estimation of parameters for the Size

Parameter	α_1	α_2	α_3	β_2	β_3
Estimate	2.22	38.77	1.76	0.89	-0.70

Prediction formulas

- Let $\mathbf{X} = (X_1, X_2, X_3) \sim h_{\boldsymbol{U}}$ a MGPD vector

Conditionnal distribution of X_3 given X_1 and X_2

$$P[X_3 \geq \ell | X_2 = x_2, X_1 = x_1] = \frac{\int_{x_3=\ell}^{\infty} 1_{\{\boldsymbol{x} \not\leq \mathbf{0}\}} \int_0^{\infty} f_{\boldsymbol{U}}(\boldsymbol{x} + \log t) \mathrm{d}t \mathrm{d}x_3}{\int_{x_3=-\infty}^{\infty} 1_{\{\boldsymbol{x} \not\leq \mathbf{0}\}} \int_0^{\infty} f_{\boldsymbol{U}}(\boldsymbol{x} + \log t) \mathrm{d}t \mathrm{d}x_3}.$$

- When $\boldsymbol{U}$ has independent components

$$P[X_3 \geq \ell | X_2 = x_2, X_1 = x_1] = \begin{cases} \dfrac{\int_0^{\infty} f_1(x_1 + \log t) f_2(x_2 + \log t)(1 - F_3(\ell + \log t)) \mathrm{d}t}{\int_0^{\infty} f_1(x_1 + \log t) f_2(x_2 + \log t) \mathrm{d}t}, & \text{if } x_1 \vee x_2 > 0 \\ \dfrac{\int_0^{\infty} f_1(x_1 + \log t) f_2(x_2 + \log t)(1 - F_3(\ell + \log t)) \mathrm{d}t}{\int_0^{\infty} f_1(x_1 + \log t) f_2(x_2 + \log t)(1 - F_3(0)) \mathrm{d}t}, & \text{if } x_1 \vee x_2 \leq 0 \end{cases}$$

(P2) Application to the data

Prediction of 2019 epidemic

In 2019, suppose we have observed the two first weeks

$$y_1 = 336 \quad \text{and} \quad y_2 = 540$$

Multiple	0.5	0.75	0.95	1	1.2
Level	864	1,297	1,642	1,729	2,075
GP	0.185	0.012	0.0023	0.0006	0.00007
Logistic	0.174	0.104	-	-	-

a) Week 3

Multiple	0.5	0.75	0.95	1	1.2
Level	4,031	6,046	7,659	8,062	9,674
GP	0.025	0.008	0.00292	0.0023	0.0009
Logistic	0.14	0.10	-	-	-

b) Size of the epidemic

(P2) Application to the data

Prediction of 2019 epidemic

In 2019, suppose we have observed the two first weeks

$$y_1 = 336 \quad \text{and} \quad y_2 = 540$$

Multiple	0.5	0.75	0.95	1	1.2
Level	864	1,297	1,642	1,729	2,075
GP	0.185	0.012	0.0023	0.0006	0.00007
Logistic	0.174	0.104	-	-	-

a) Week 3

Multiple	0.5	0.75	0.95	1	1.2
Level	4,031	6,046	7,659	8,062	9,674
GP	0.025	0.008	0.00292	0.0023	0.0009
Logistic	0.14	0.10	-	-	-

b) Size of the epidemic

In fact, $y_3 = 599$ and $s_3 = 1,505$

(P3) Detection of abnormal epidemics

- Take advantage of the GP method for early dectetion of "unusual", potentially very dangourous, epidemics
- **Abnormal** = abnormal w.r.t. the fitted GP model
- **Anomaly** = unusual observation given that the ILI rates of at least one of Weeks 1, 2 or 3 is large

  $\hookrightarrow$ A very large value of the estimated GP negative log-likelihood of the observed rates during the first 3 weeks

  $\Rightarrow$ something unusual is happening
- GP negative log-likelihood

$$-\log h_{\boldsymbol{U}}(\boldsymbol{x}) = \log \mathbb{E}[\mathrm{e}^{\max \boldsymbol{U}}] - \log \int_0^\infty f_{\boldsymbol{U}}(\boldsymbol{x} + \log t)\,\mathrm{d}t,$$

  for $\boldsymbol{x} \in \mathbb{R}^3$ such that $\boldsymbol{x} \not\leq \boldsymbol{0}$.
- From 1,500 simulated datasets, estimate significance levels

(P3) Detection of abnormal epidemics

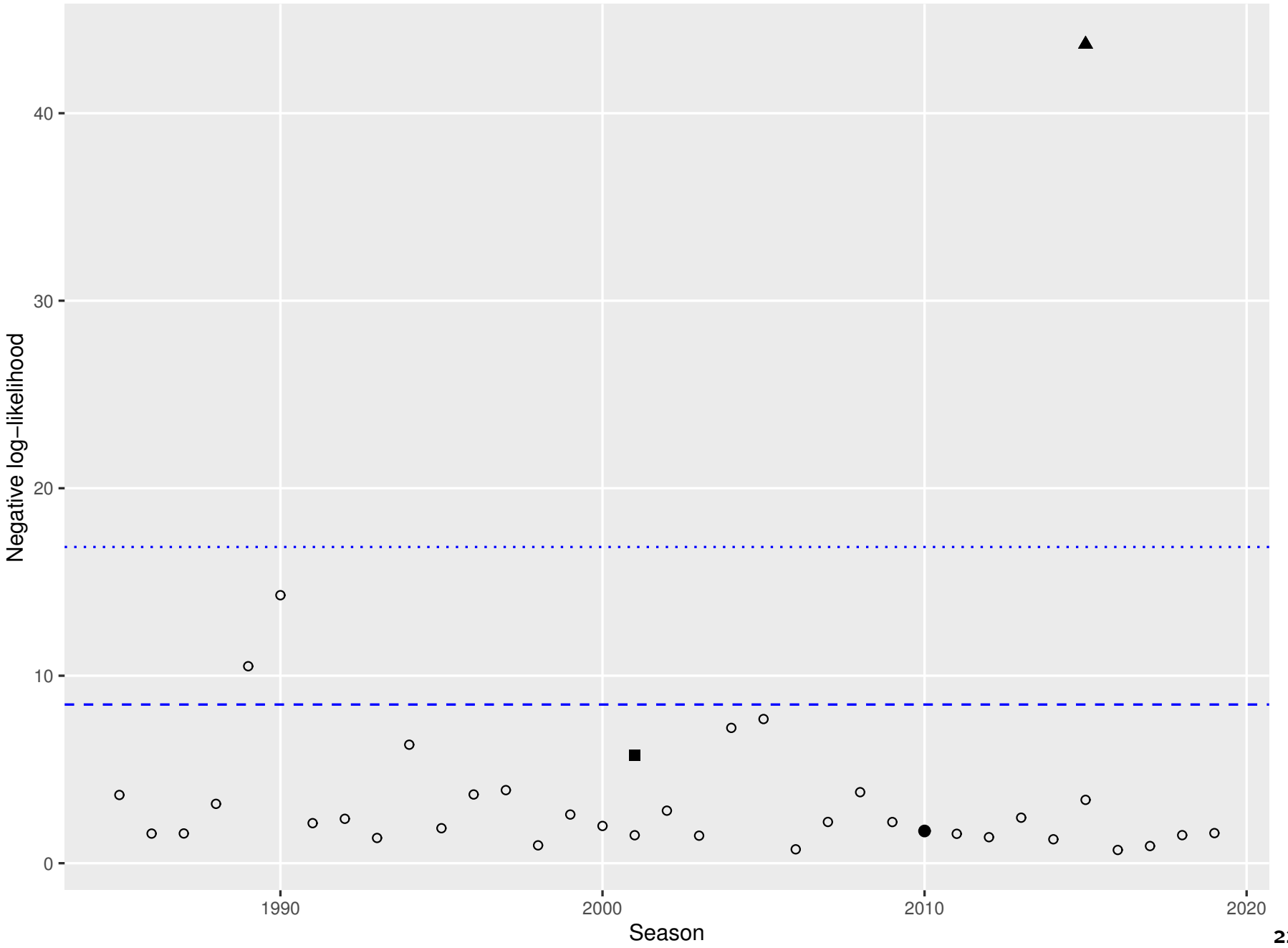
Significance level	10%	5%	1%	0.1%
Cut-off	4.87	5.75	8.46	16.86

There is a fine line between extreme and abnormal epidemics

Add two new epidemics (extreme and abnormal)

1. Simulate from the Gumbel an epidemic with the ILI rate of Week 3 equal to the 0.99 quantile of the simulated rates for Week 3
2. Multiply the rates of Weeks 1 and 2 by 2 and the rate of Week 2 by 0.1

(P3) Detection of abnormal epidemics



What about Covid-19?

- Difficult to apply to new viruses
  → Lack of data and historical information
  → Very few knowledge, no hindsight
  → Pandemics have different behaviours compare to seasonal epidemics

- Extreme Value Theory ⇒ Inference outside of the range of the sample
  → Application in Public Health domain can help
  → Predict the need in masks
  → Predict of overcrowding in hospitals

What about Covid-19?

- Difficult to apply to new viruses
  → Lack of data and historical information
  → Very few knowledge, no hindsight
  → Pandemics have different behaviours compare to seasonal epidemics

- Extreme Value Theory ⇒ Inference outside of the range of the sample
  → Application in Public Health domain can help
  → Predict the need in masks
  → Predict of overcrowding in hospitals

Thank you for your attention!