Méthodes statistiques pour l'épidémiologie

Felix Cheysson

Sorbonne Université, LPSM (UMR 8001).

9èmes Journées YSP
Jeudi 28 janvier 2021

Outline

1. Overview of the statistical methods in epidemiology
   - You know something, John Snow
   - Diversity of biostatistics
   - The cell, basic unit of life
2. Modelling the risk of death for Covid-19 patients
   - Logistic regression
   - Extending the logistic regression
   - Moment estimator from censored data

Outline

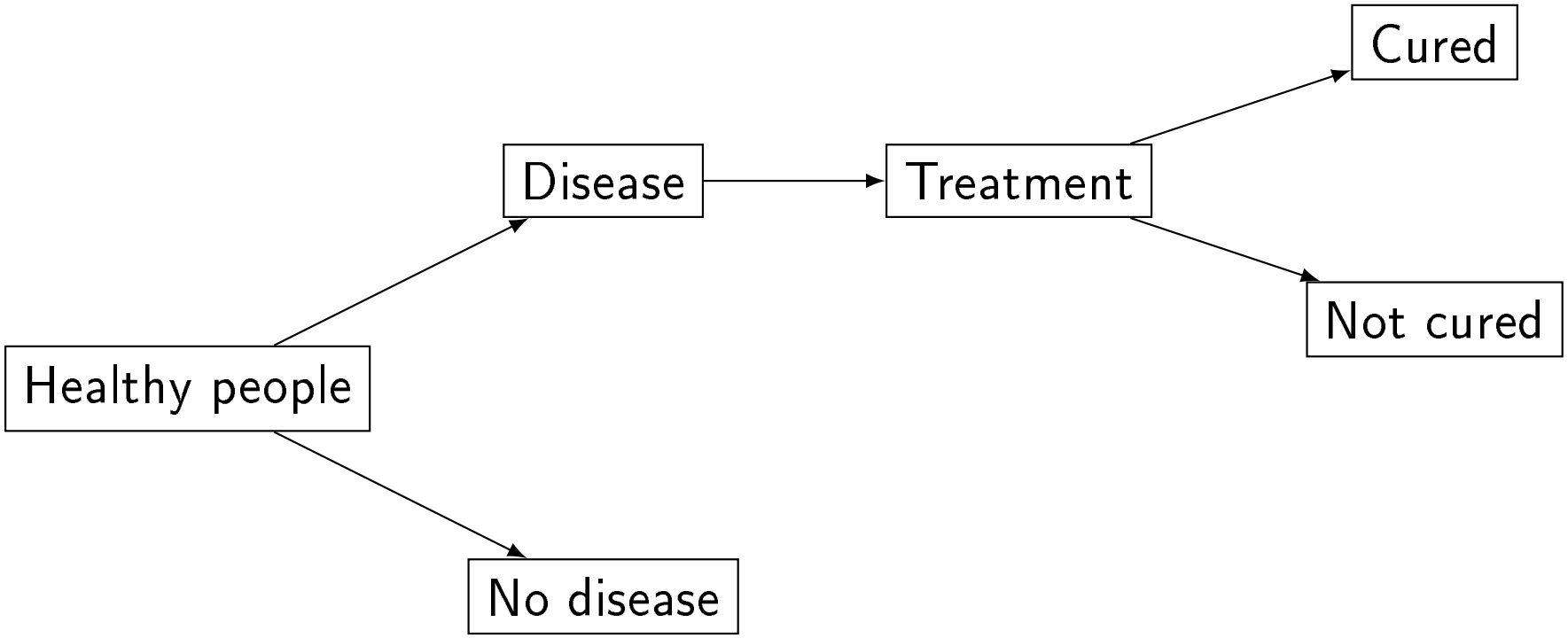
John Snow[1], father of field epidemiology

- Studied cholera outbreaks to discover their cause and to prevent them.
- Descriptive epidemiology to hypothesis generation and testing to application.
- The Broad Street pump, 1854 (Snow, 1936).

[1] Not the one in Game of Thrones

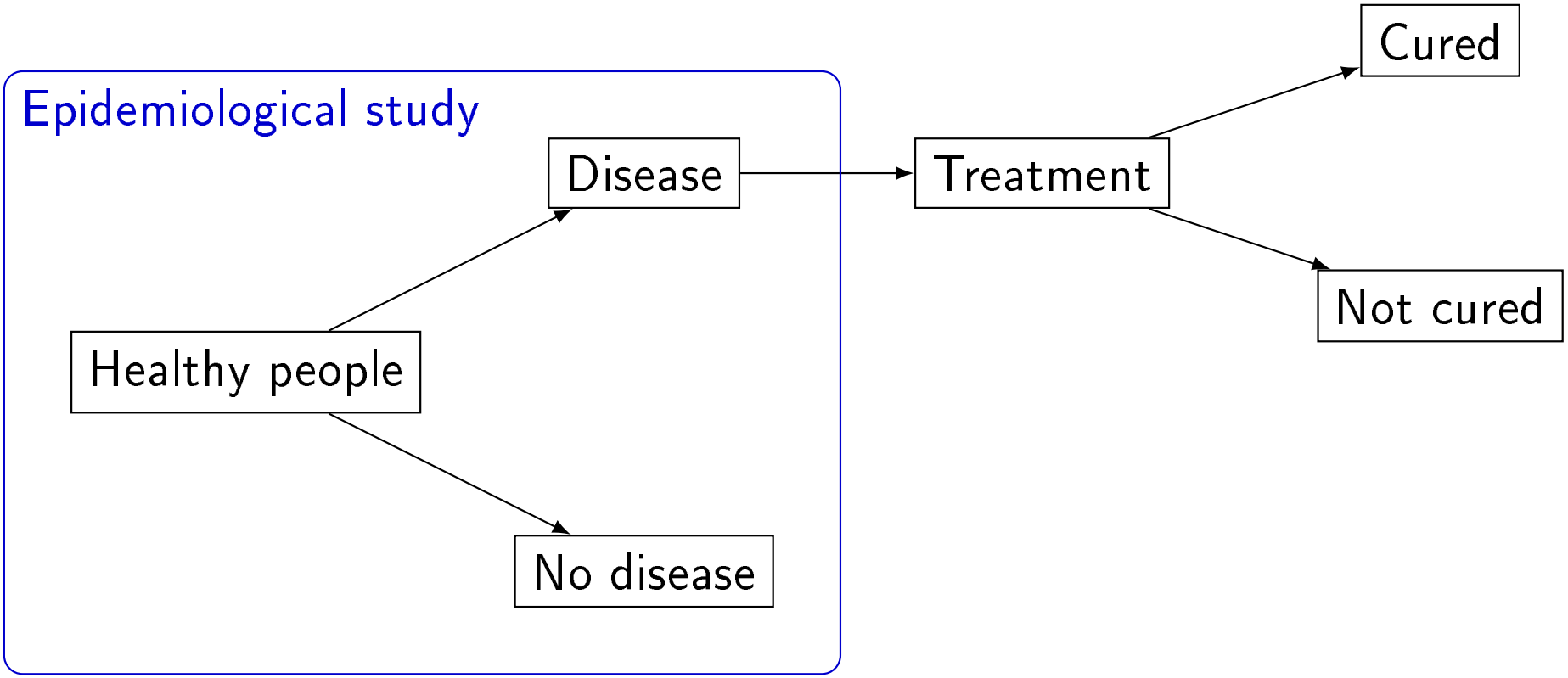
Studies in public health

Epidemiology: study and analysis of the distribution (who, when, and where), patterns and determinants of health and disease conditions in defined populations (Wikipedia).



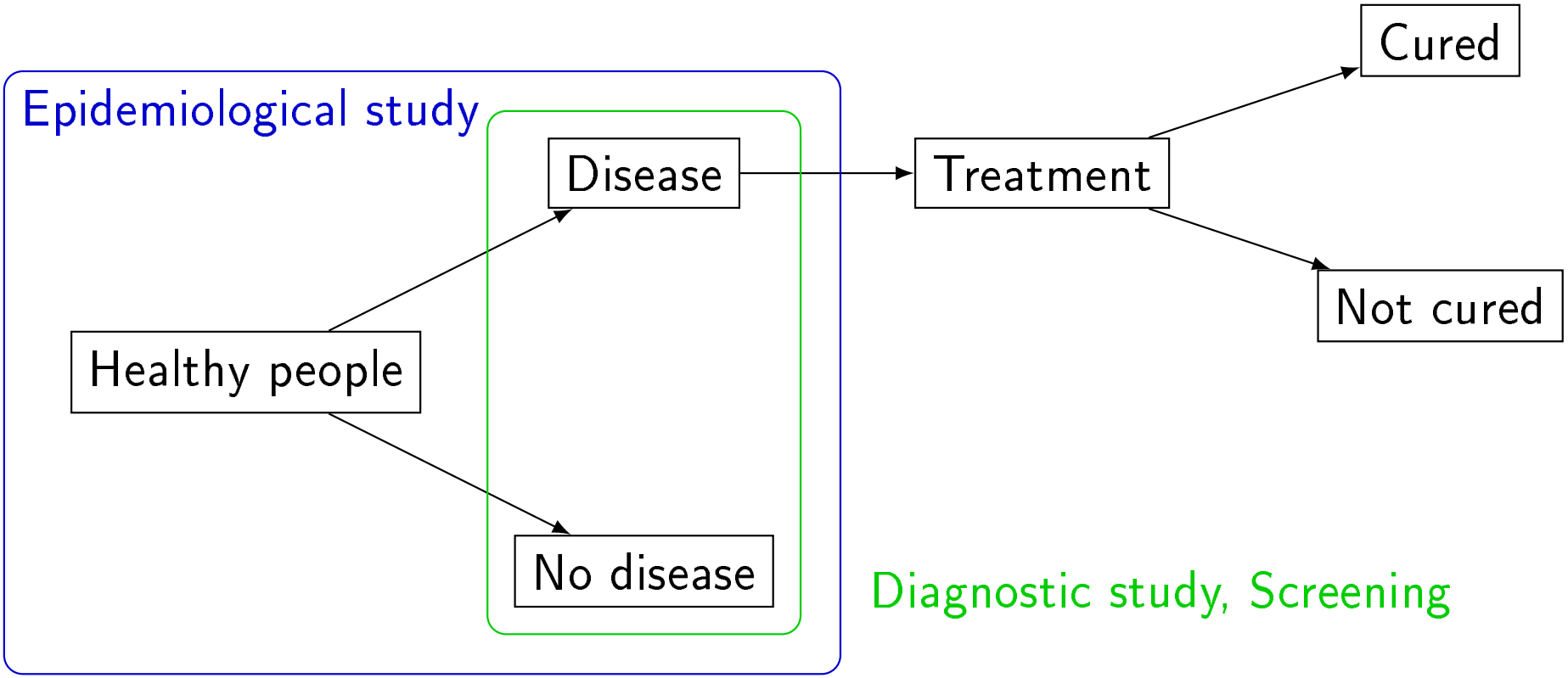
Studies in public health

Epidemiology: study and analysis of the distribution (who, when, and where), patterns and determinants of health and disease conditions in defined populations (Wikipedia).



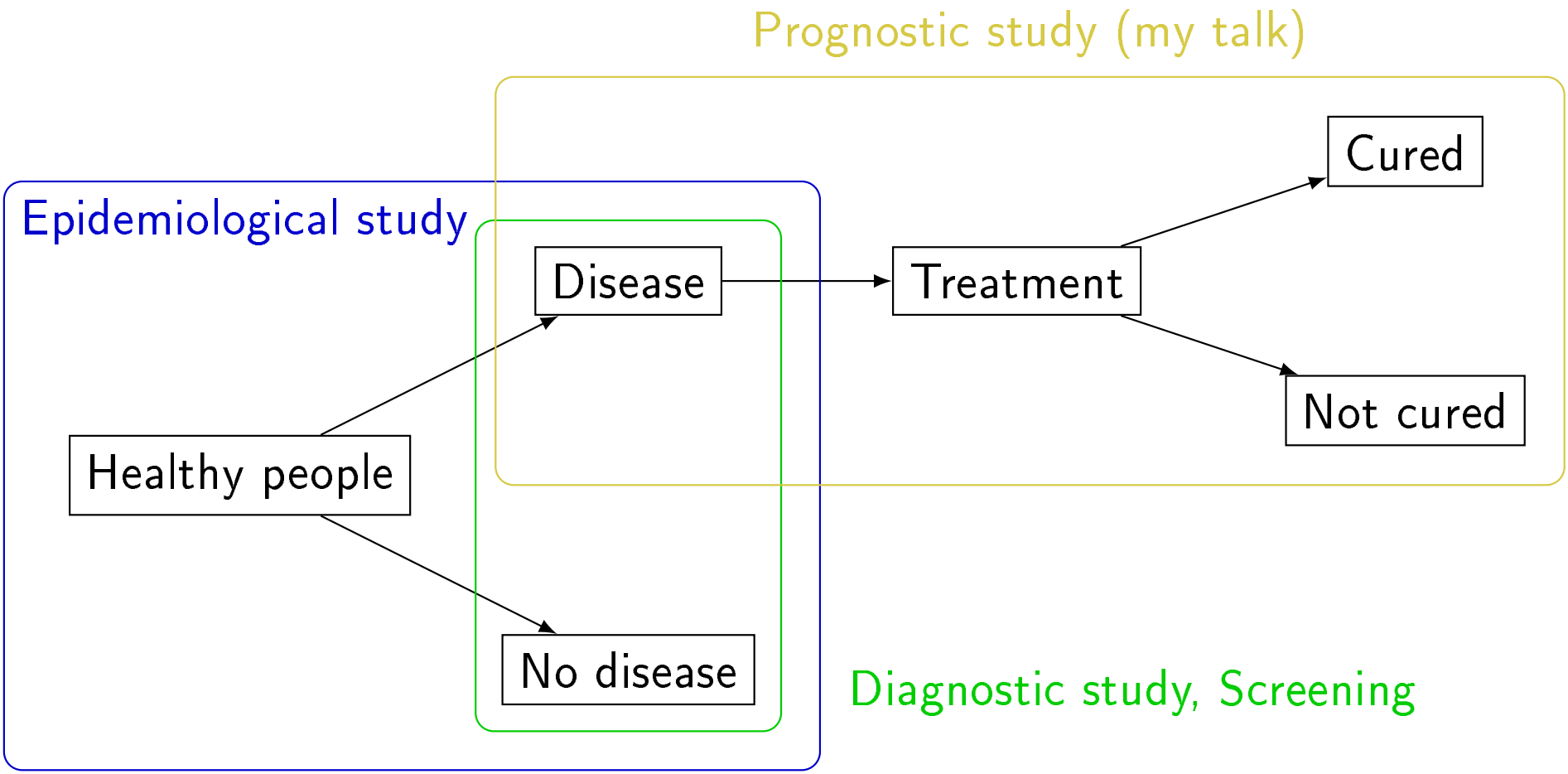
Studies in public health

Epidemiology: study and analysis of the distribution (who, when, and where), patterns and determinants of health and disease conditions in defined populations (Wikipedia).



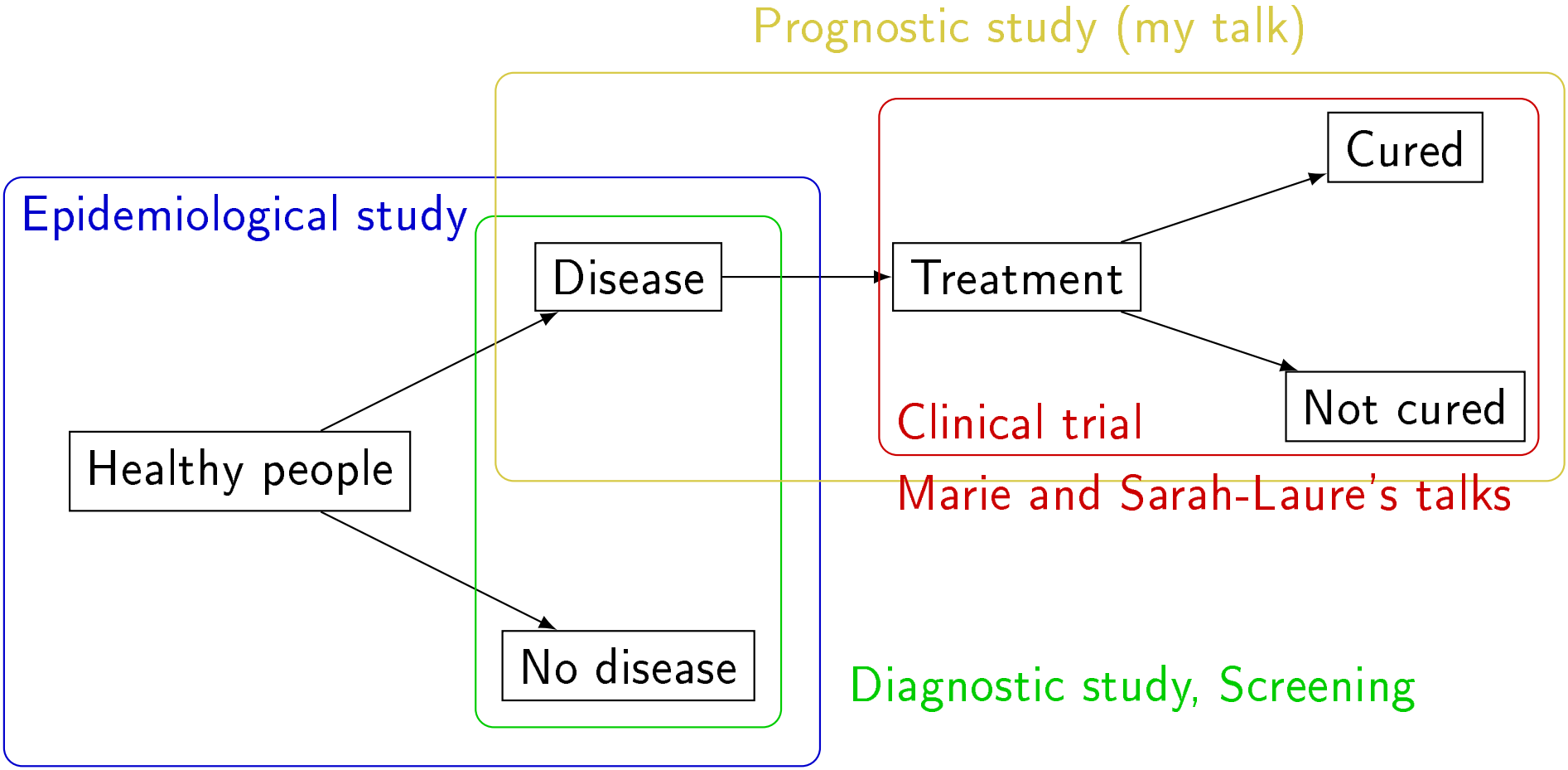
Studies in public health

Epidemiology: study and analysis of the distribution (who, when, and where), patterns and determinants of health and disease conditions in defined populations (Wikipedia).



Studies in public health

Epidemiology: study and analysis of the distribution (who, when, and where), patterns and determinants of health and disease conditions in defined populations (Wikipedia).



Statistical methods and models (non exhaustive)

Epidemiological studies

- Descriptive statistics and moment estimators (Bard et al., 2005);
- Generalised linear models and mixed models (all talks);
- Point processes (Meyer, Elias, and Höhle, 2012).

Prognostic studies:

- Predictive models, machine learning;
- Survival analysis.

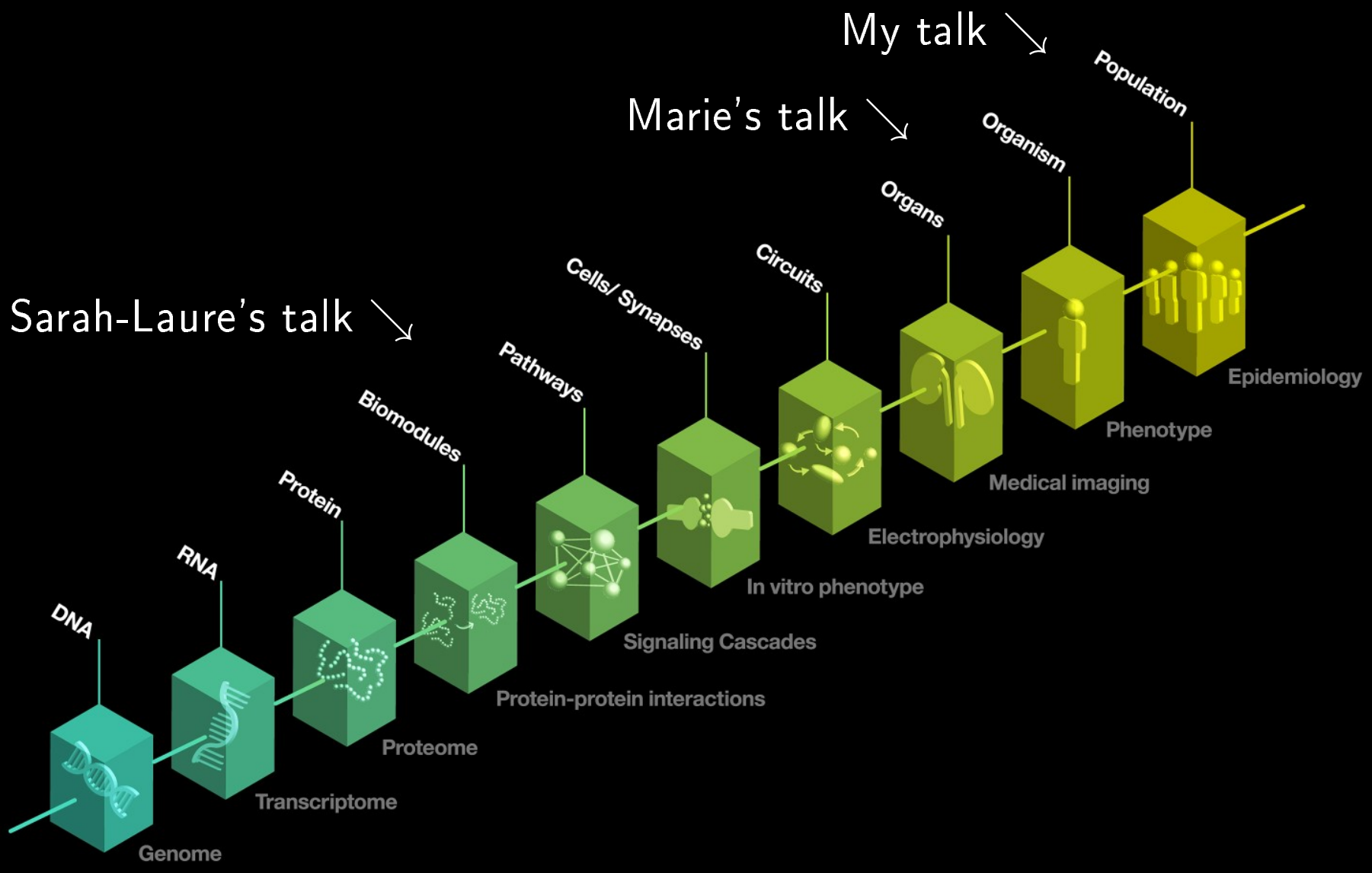
Clinical trials:

- Design of experiment;
- Hypothesis tests (Marie's talk).

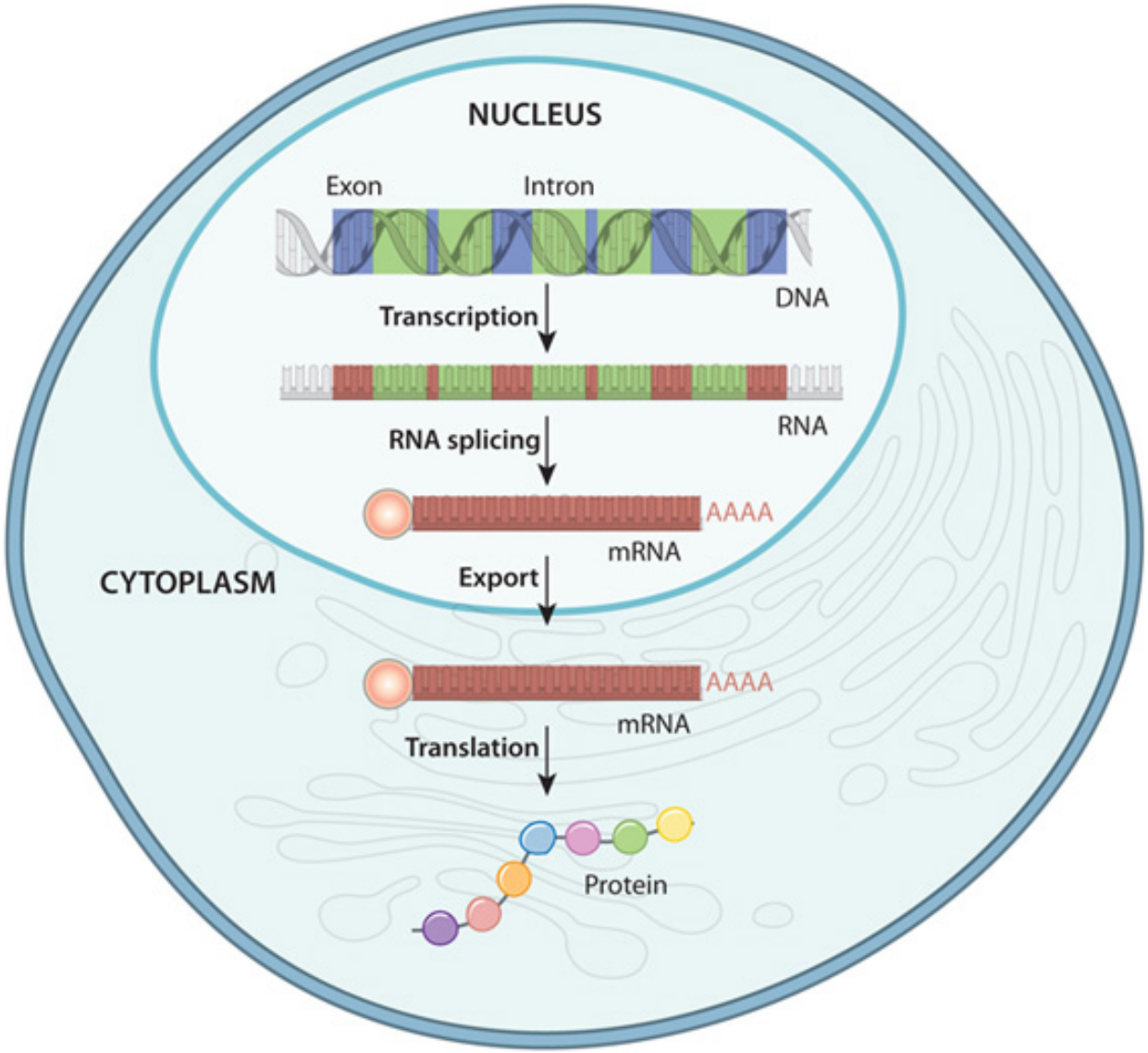
Genomics:

- Dimensionality reduction (ACP, factor analysis) (Sarah-Laure's talk);
- Penalised regressions (Grandclaudon et al., 2019).

Multiple scales of biology



From DNA to protein



Outline

Modelling motivation

Covid-19 Dataset from SI-VIC database: all hospitalisation for Covid-19 patients in AP-HP hospitals.

dt.first	dt.last	outcome	sex	age	hospital
2020-03-17	2020-04-05	rad	F	45	Robert Debré
2020-03-14	2020-03-25	rad	F	29	Robert Debré
2020-03-18	2020-03-29	dc	H	80	St Antoine
2020-03-11	2020-03-15	dc	H	62	St Louis
2020-03-04	2020-03-09	dc	F	72	Pitié Salpétrière
2020-03-16	2020-03-20	dc	H	92	Raymond Poincaré

Motivation: We wish to model the probability of an event occurring, *e.g.* the risk of death of a patient hospitalised for Covid-19.

Logistic regression

For individual i, let X_i denote their age and Y_i the outcome of hospitalisation:

$$\begin{aligned} Y_i &= 1, && \text{if the individual } i \text{ dies,} \\ Y_i &= 0, && \text{if the individual } i \text{ lives.} \end{aligned}$$

From an *i.i.d.* sample $\big((x_1, y_1), \ldots, (x_n, y_n)\big)$, we wish to explain the risk of death as a function of the age of the individual:

$$p_i = \mathbb{P}(Y_i = 1 | X_i = x_i).$$

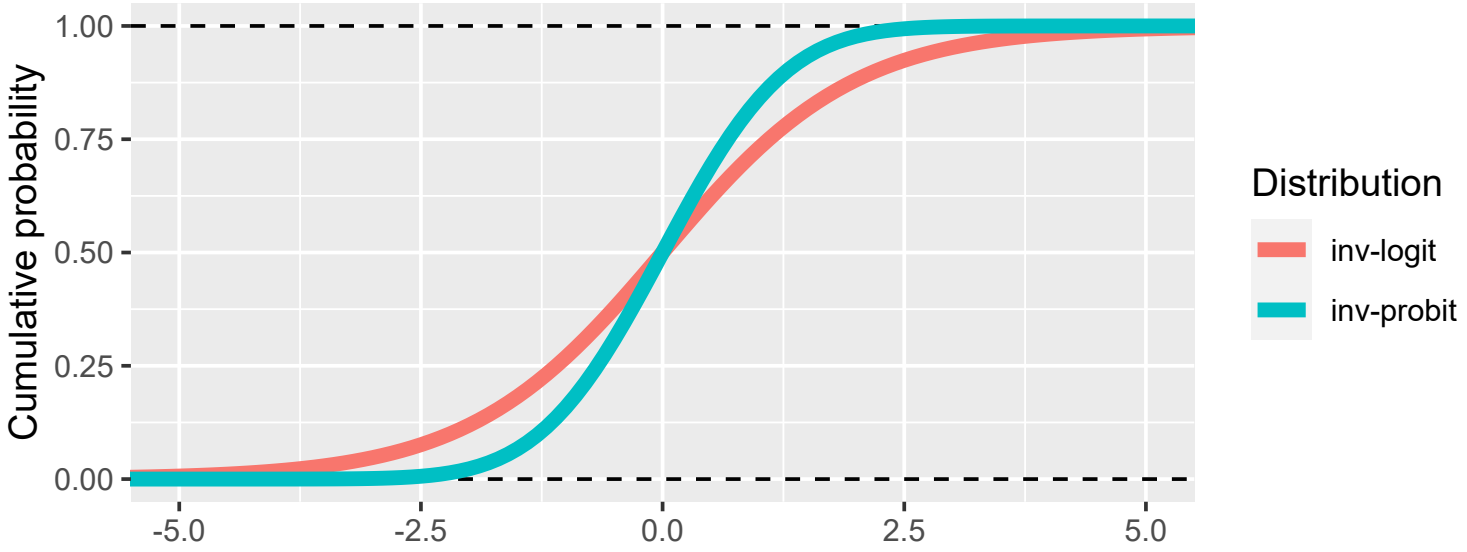
We model the outcome with a logistic regression:

$$\begin{aligned} Y_i &\overset{\textit{ind.}}{\sim} B(p_i), \\ g(p_i) &= \beta_0 + \beta_1 x_i. \end{aligned}$$

Choice of the link function g

- g must be chosen as a map from $(0, 1)$ to $\mathbb{R}$.
- Two usual choices:
  - The *probit* function: $\text{probit}(p_i) = \Phi^{-1}(p_i)$, where $\Phi(x)$ is the CDF of the normal distribution.
  - The *logit* function: $\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$.
- The logit function can be easily interpreted in terms of *odds-ratio*:

$$\text{logit}(p_1) - \text{logit}(p_2) = \log\left(\frac{p_1/(1-p_1)}{p_2/(1-p_2)}\right).$$

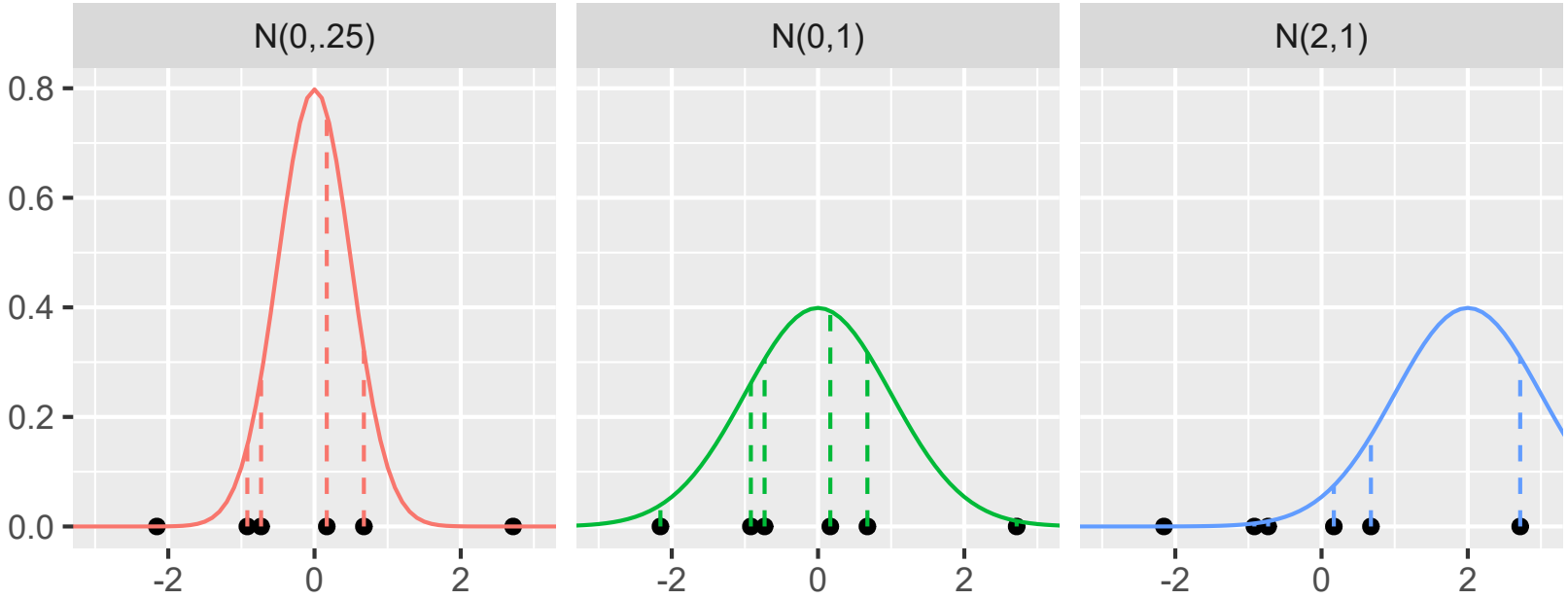


Likelihood function

General case: Suppose that Y_i has density $f_\theta(y)$. Then the function $L_i(\theta) = f_\theta(Y_i)$ w.r.t. θ is called the likelihood function of θ.
Objective: Maximise $L_i(\theta)$, equivalently $l_i(\theta) = \log L_i(\theta)$, w.r.t. $\theta \in \Omega$,

$$\hat{\theta} = \arg\max_{\theta \in \Omega} L_i(\theta) = \arg\max_{\theta \in \Omega} l_i(\theta).$$

▷ $\hat{\theta}$ is called the maximum likelihood estimator of θ.



Consistency and asymptotic properties

- Suppose that the data $(Y_1, \ldots, Y_n)$ is generated from distribution $f_{\theta_0}(y)$ with true parameter θ_0.
- The log-likelihood of the model is written $l_n(\theta) = \frac{1}{n}\sum_{i=1}^n l_i(\theta)$.
- For the **logistic regression**,

$$l_n(\theta) = \frac{1}{n}\sum_{i=1}^n Y_i \log p_i + (1 - Y_i)\log(1 - p_i).$$

- Define $\hat{\theta}_n$ as the maximum likelihood estimator of θ.

Theorem: Under regularity conditions, $\hat{\theta}_n$ is consistent, *i.e.* $\hat{\theta}_n \xrightarrow{P} \theta_0$, and is asymptotically normal:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to \mathcal{N}\left(0, \frac{1}{I(\theta_0)}\right),$$

where $I(\theta_0) = \mathbb{E}_{\theta_0}\left[\left(\frac{\partial}{\partial\theta}\log f_\theta(Y)\,|_{\theta=\theta_0}\right)^2\right]$ is called the Fisher information.

Back to the data

Asymptotic confidence interval

Using the asymptotic normality of the MLE $\hat{\theta}_n$, we build approximate confidence interval for θ_0 for n large:

$$\mathrm{IC}_{1-\alpha}(\theta_0) \approx \left[\hat{\theta}_n + \frac{u_{\alpha/2}}{\sqrt{nI(\theta_0)}}; \hat{\theta}_n + \frac{u_{1-\alpha/2}}{\sqrt{nI(\theta_0)}}\right],$$

with u_a the quantile of order a of the normal distribution.

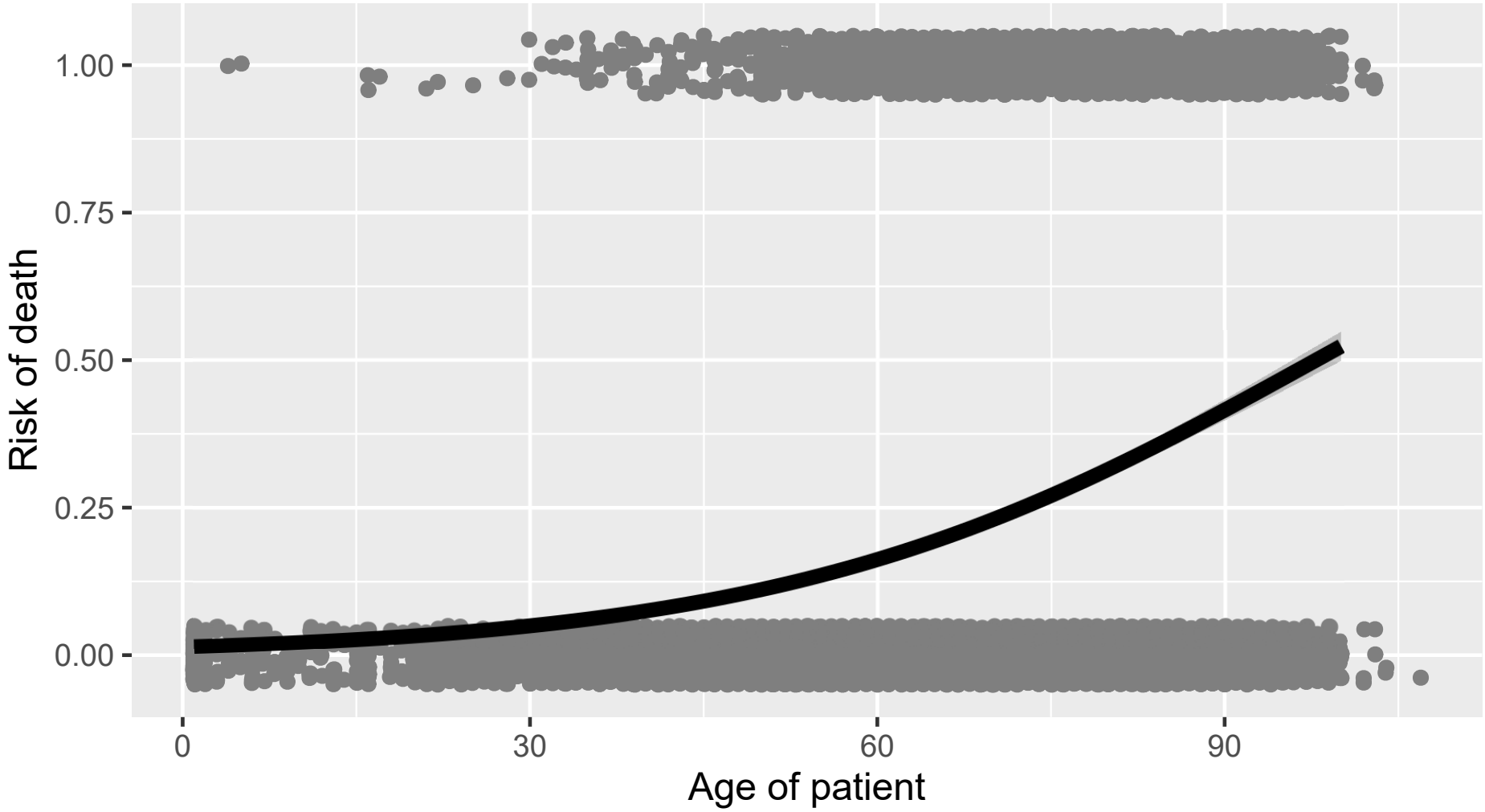
Using R, we find:

$$\mathrm{IC}_{95\%}(\beta_1) = [0.040; 0.046],$$

or, as an odds-ratio:

$$\mathrm{IC}_{95\%}\left(e^{\beta_1}\right) = [1.041; 1.048].$$

Predicting the risk of death



What about categorical variables?

For individual i, let Z_i denote their sex. We now wish to explain the risk of death as a function of the age and sex of the individual:

$$p_i = (Y_i = 1 | X_i = x_i, Z_i = z_i).$$

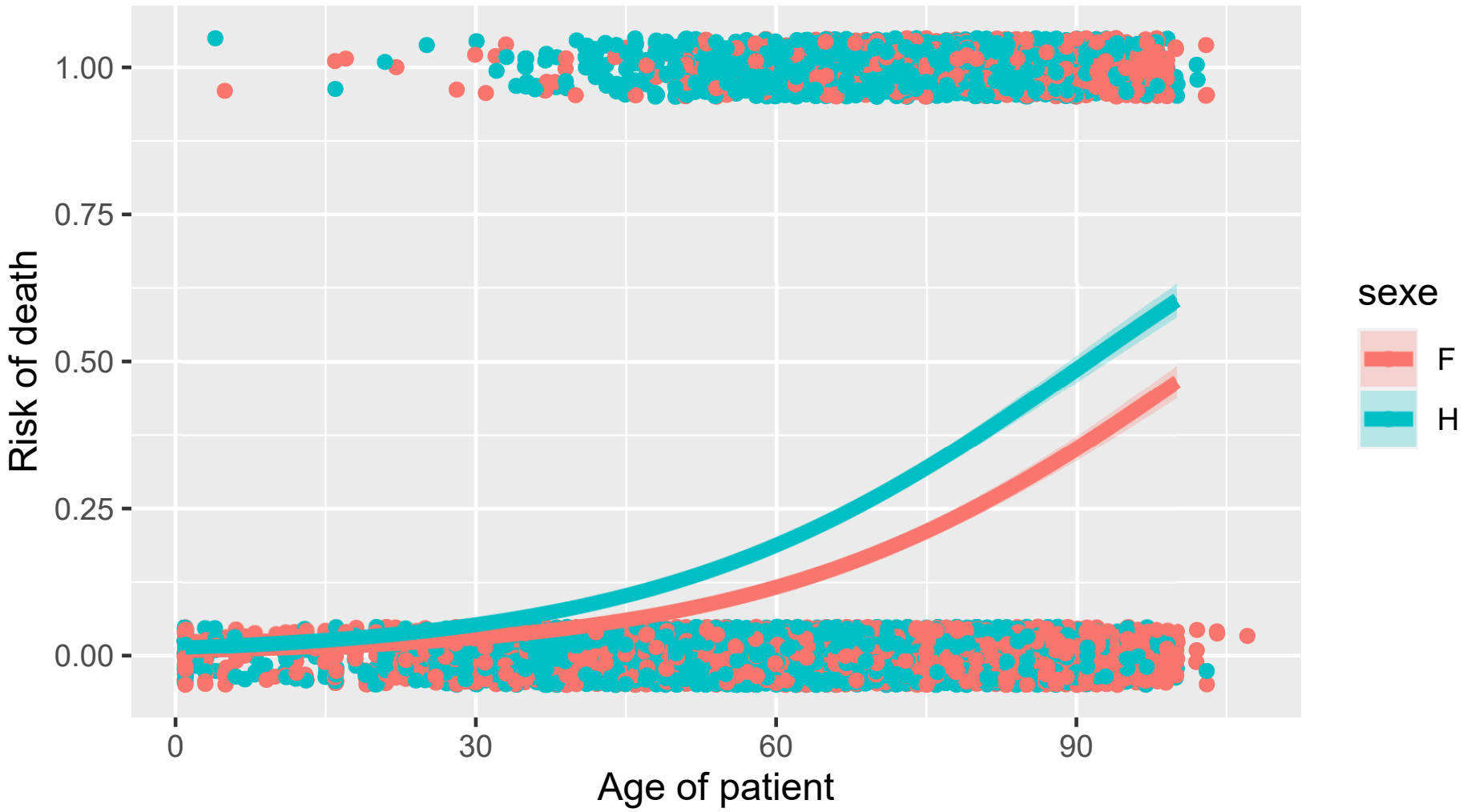
The logistic regression can be extended accordingly:

$$\begin{aligned} Y_i &\overset{ind.}{\sim} B(p_i), \\ \text{logit}(p_i) &= \beta_0 + \beta_1 x_i + \beta_2 \mathbb{1}_{\{z_i = \text{“}H\text{”}\}}. \end{aligned}$$

Using R, we find:

$$\text{IC}_{95\%}\left(e^{\beta_2}\right) = [1.570; 1.946].$$

Predicting the risk of death (w.r.t. sex)



What about hierarchical data?

For individual i, let H_i denote the hospital in which they were treated. We now wish to control for hospital differences in the model.

Problem: Observations Y_i are no longer independent since the outcomes for patients from the same hospital are correlated.

We can model the outcome with a **mixed model**:

$$Y_i|\gamma_{h_i} \overset{ind.}{\sim} B(p_i),$$
$$\text{logit}(p_i) = \beta_0 + \beta_1 x_i + \gamma_{h_i},$$

where $\gamma_h \overset{i.i.d.}{\sim} \mathcal{L}(\vartheta)$, usually $\mathcal{N}(\nu, \varsigma^2)$.

Likelihood: Using Bayes' theorem, $f_{\theta,\vartheta}(Y, \gamma_h) = f_\theta(Y|\gamma_h) f_\vartheta(\gamma_h)$.

Covariance of Y_i

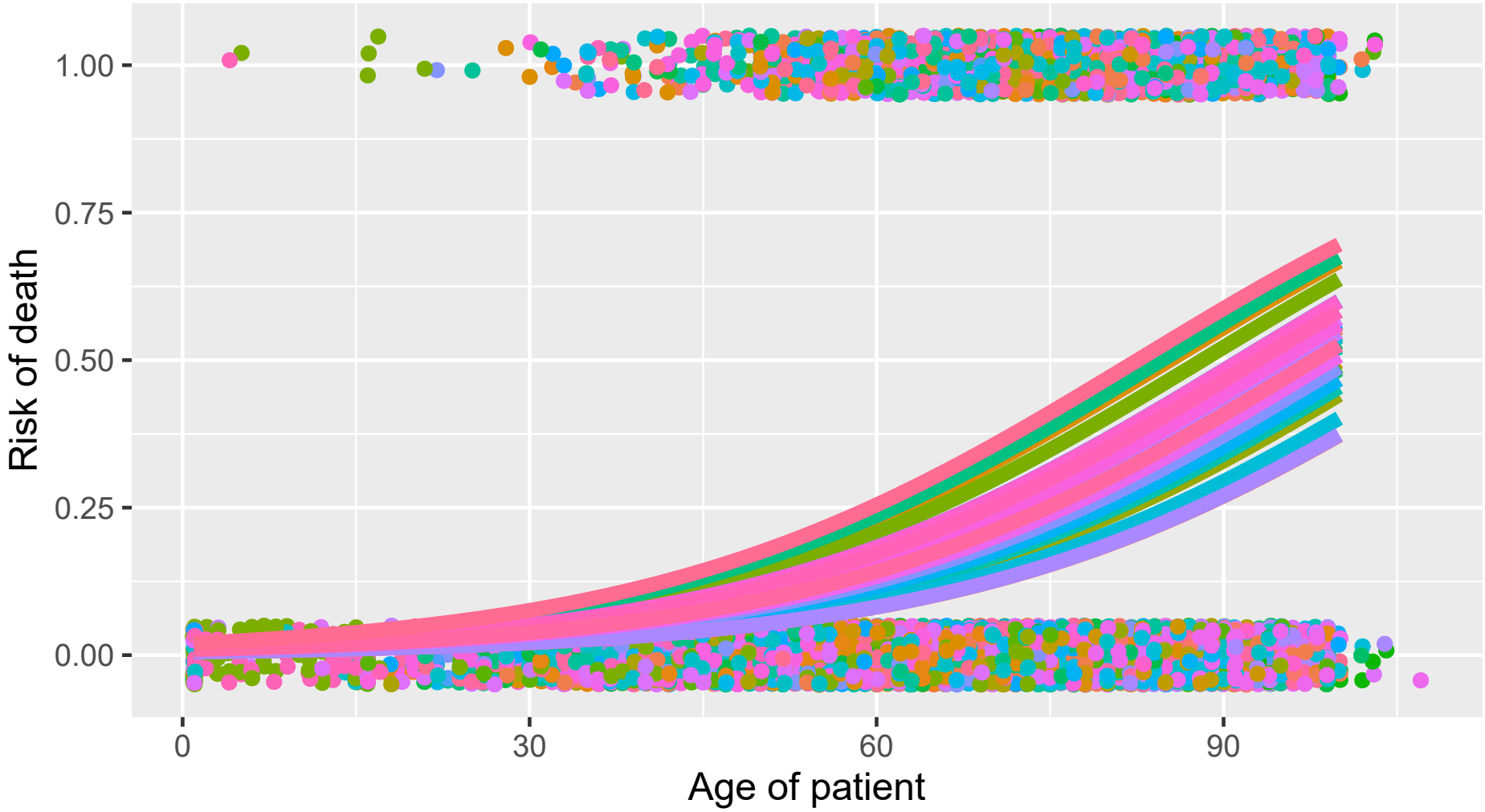
Using the law of total covariance (can be obtained using conditional expectation):

$$\mathrm{Cov}(Y_i, Y_j) = \mathbb{E}\big[\mathrm{Cov}(Y_i, Y_j|\gamma_h)\big] + \mathrm{Cov}\big(\mathbb{E}[Y_i|\gamma_h], \mathbb{E}[Y_j|\gamma_h]\big).$$

- We have $\mathrm{Cov}(Y_i, Y_j|\gamma_h) = 0$ from the model definition;
- and $\mathbb{E}[Y_i|\gamma_h] = \mathrm{logit}^{-1}(\beta_0 + \beta_1 x_i + \gamma_{h_i})$;
- therefore,

$$\mathrm{Cov}(Y_i, Y_j) \begin{cases} = 0, & \text{if } h_i \neq h_j, \\ \neq 0, & \text{if } h_i = h_j. \end{cases}$$

Predicting the risk of death (w.r.t. hospitals)



The Bayesian framework

Idea: Assume model on both the parameters and the data,

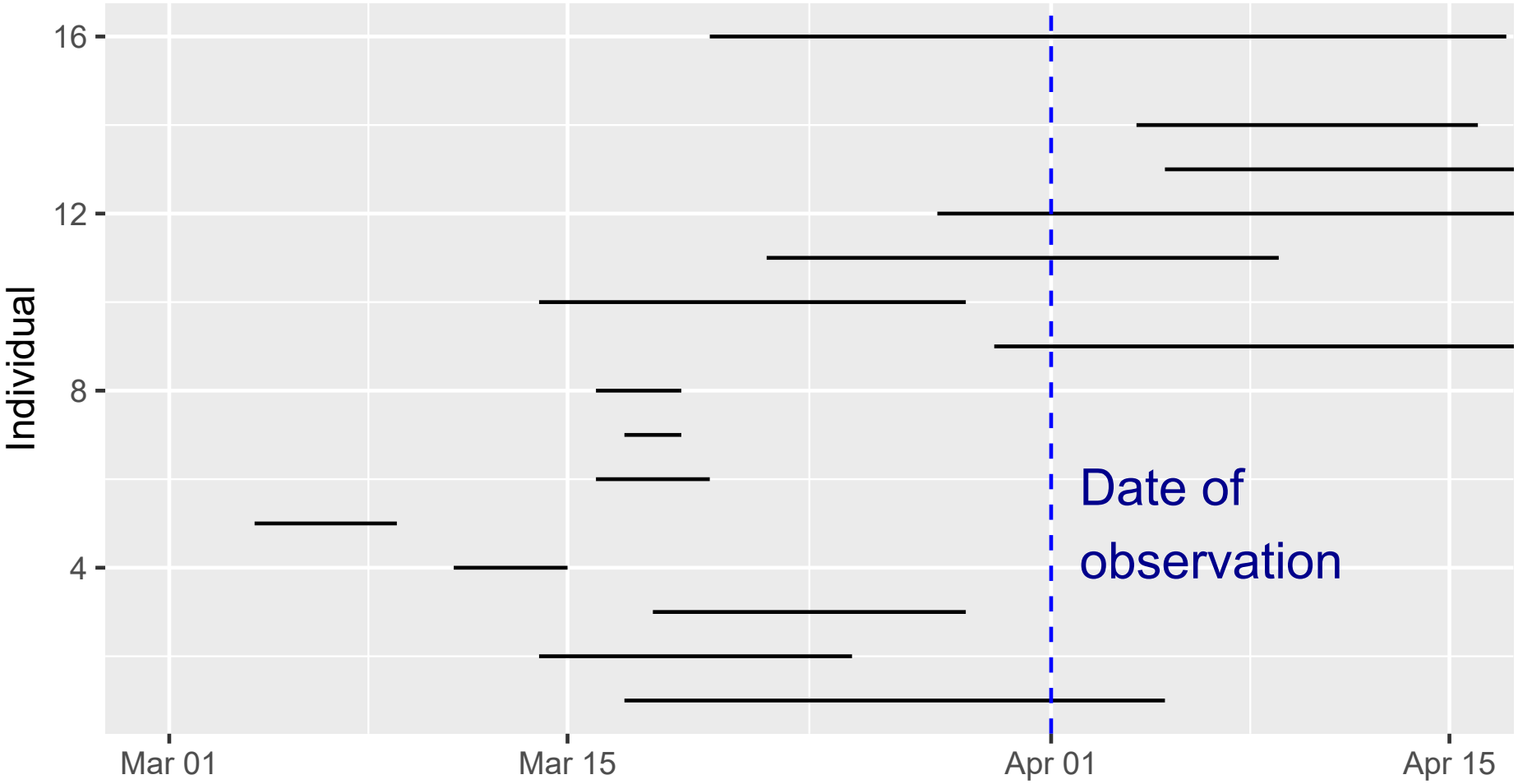
$$f(\theta|Y) \;=\; \frac{f(Y|\theta)f(\theta)}{f(Y)} \;=\; \frac{f(Y|\theta)f(\theta)}{\int f(Y|\theta)f(\theta)\mathrm{d}\theta} \;\propto\; f(Y|\theta)f(\theta),$$

where $f(Y|\theta)$ is the likelihood, and $f(\theta)$ is called the prior distribution of θ.

When to use the Bayesian framework:

- The prior distribution can be tailored by expert knowledge, to add *a priori* information to the estimation.
  Remark: The prior distribution can be related to a regularisation term.
- Use of MCMC algorithm to sample from the *posterior distribution* $f(\theta|Y)$ readily available (WinBUGS, OpenBUGS, JAGS, Stan, ...)
  Remark: These softwares handle the decomposition of hierarchical models through successive applications of Bayes' theorem.

Censored data



Some notations

For an individual i, denote by

- E_i: day of hospitalisation;
- T_i: length of hospitalisation;
- U_i: outcome of the hospitalisation (1 = cured, 2 = dead);

We are interested in the quantity $\pi = \mathbb{P}(U = 2)$.

Problem: data are right-censored, and we only observe, at date x,

$$\begin{cases} C_i &= x - E_i, \\ Y_i &= \min(T_i, C_i), \\ \delta_i &= 1\{T_i \leq C_i\}, \\ Z_i &= \delta_i U_i. \end{cases}$$

Estimator in the case of censoring data

Assume that (T_i, U_i) is independent from C_i. Define the survival function of the censoring process

$$S_C(t) = \mathbb{P}(C_1 \geq t).$$

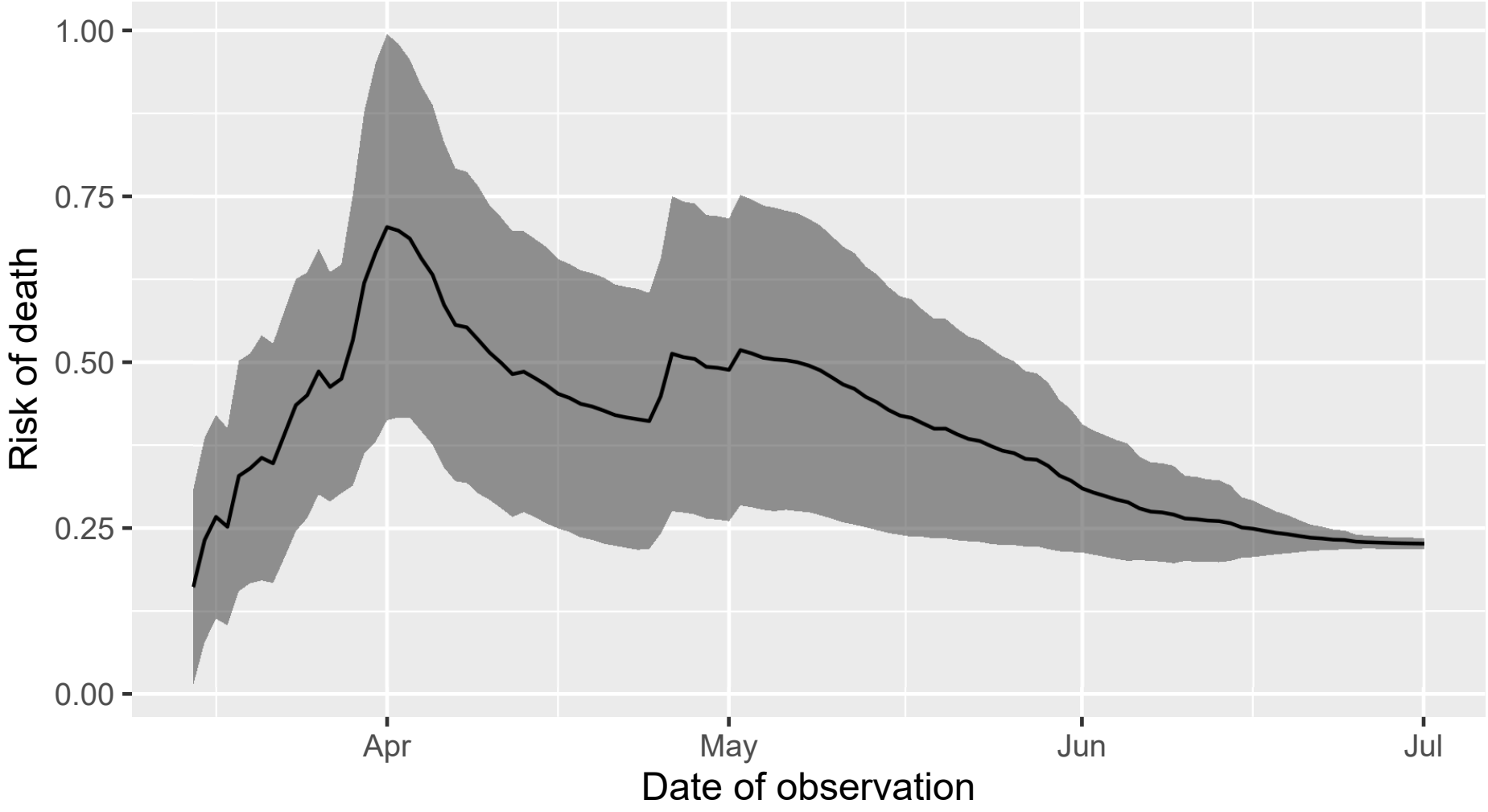
Since

$$\begin{aligned}
\mathbb{E}\left[\frac{\delta_1 1\{U_i = 2\}}{S_C(Y_i)}\right] &= \mathbb{E}\left[\frac{1\{U_i = 2\}}{S_C(T_i)}\mathbb{E}[1\{T_i \leq C_i\} \mid T_i, U_i]\right] \\
&= \mathbb{E}\left[\frac{1\{U_i = 2\}}{S_C(T_i)} S_C(T_i)\right] \\
&= \mathbb{E}\left[1\{U_i = 2\}\right] \\
&= \mathbb{P}(U_i = 2) = \pi,
\end{aligned}$$

a natural estimator of π, at date of observation x, is the quantity

$$\hat{\pi}_x = \frac{1}{n_x} \sum_{i=1}^{n_x} \frac{\delta_{i,x} 1\{U_i = 2\}}{\hat{S}_{C,x}(Y_{i,x})}.$$

Online estimation of risk of death



For Further Reading I

- Bard, Denis et al. (2005). "Risque attribuable". In: *Cancer - Approch. méthodologique du lien avec l'environnement*. Les éditions Inserm, pp. 69–92. isbn: 2-85598-844-6.
- Grandclaudon, Maximilien et al. (2019). "A Quantitative Multivariate Model of Human Dendritic Cell-T Helper Cell Communication". In: *Cell* 179.2, pp. 432–447.
- Meyer, Sebastian, Johannes Elias, and Michael Höhle (2012). "A Space-Time Conditional Intensity Model for Invasive Meningococcal Disease Occurrence". In: *Biometrics* 68.2, pp. 607–616. issn: 0006341X. doi: 10.1111/j.1541-0420.2011.01684.x. arXiv: 1508.05740.
- Snow, John (1936). *Snow on cholera. London: Humphrey Milford*.

Some notations

- Suppose that the data $(Y_1, \ldots, Y_n)$ is generated from distribution $f_{\theta_0}(y)$ with true parameter θ_0.
- The log-likelihood of the model is written

$$l_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} l_i(\theta).$$

- For the **logistic regression**,

$$l_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} Y_i \log p_i + (1 - Y_i) \log(1 - p_i).$$

- Define $\hat{\theta}_n$ as the maximum likelihood estimator of θ.

Consistency of the likelihood

Define, for the expected value of $l_1(\theta)$:

$$l(\theta) = \mathbb{E}_{\theta_0}[l_1(\theta)] = \int (\log f_\theta(y)) f_{\theta_0}(y) \mathrm{d}y.$$

Lemma: For any θ,

$$l(\theta) \leq l(\theta_0).$$

If the model is identifiable, then the inequality is strict for $\theta \neq \theta_0$.

Idea of the proof: Remark that the difference

$$l(\theta_0) - l(\theta) = \mathbb{E}_{\theta_0} \log \frac{f_{\theta_0}(Y)}{f_\theta(Y)}$$

is a Kullback-Leibler divergence. Show that it is non-negative (*e.g.* using Jensen's inequality).

Consistency of the likelihood (cont'd)

Define, for the expected value of $l_1(\theta)$:

$$l(\theta) = \mathbb{E}_{\theta_0}[l_1(\theta)] = \int (\log f_\theta(y)) f_{\theta_0}(y) \mathrm{d}y.$$

Theorem: If $l_n(\theta)$ is continuous and has a unique maximum, then $\hat{\theta}_n$ is consistent, *i.e.* $\hat{\theta}_n \xrightarrow{P} \theta_0$.

Idea of the proof: We have the following assertions:

- $\hat{\theta}_n$ is the maximiser of $l_n(\theta)$ (by definition);
- θ_0 is the maximiser of $l(\theta)$ (by lemma);
- $\forall \theta$, $l_n(\theta) \xrightarrow{P} l(\theta)$ (by WLLN).

Fisher information

Define, for a log-likelihood $l(\theta) = f_\theta(y)$, the **Fisher information** function by

$$I(\theta) = \mathbb{E}_\theta\big[(l'(\theta))^2\big] = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f_\theta(Y)\right)^2\right].$$

Lemma: We have the following:

$$I(\theta) = \mathrm{Var}_\theta\big(l'(\theta)\big), \qquad \text{and } I(\theta) = -\mathbb{E}_\theta\big[l''(\theta)\big].$$

Idea of the proof: We have, by swapping the derivative and the integral:

$$\int \frac{\partial}{\partial\theta} f_\theta(y)\mathrm{d}y = \frac{\partial}{\partial\theta}\int f_\theta(y)\mathrm{d}y = 0,$$

and

$$\int \frac{\partial^2}{\partial^2\theta} f_\theta(y)\mathrm{d}y = \frac{\partial^2}{\partial^2\theta}\int f_\theta(y)\mathrm{d}y = 0.$$

Asymptotic normality

Theorem: Under regularity conditions, we have that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to \mathcal{N}\left(0, \frac{1}{I(\theta_0)}\right).$$

Idea of the proof: A Taylor expansion of $l_n'(\hat{\theta}_n)$ around θ_0 gives:

$$0 = l_n'(\hat{\theta}_n) = l_n'(\theta_0) + (\hat{\theta}_n - \theta_0) l_n''(\theta_n^*),$$

for some θ_n^* between θ_0 and $\hat{\theta}_n$.
Therefore,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\frac{\sqrt{n}\, l_n'(\theta_0)}{l_n''(\theta_n^*)}.$$

Asymptotic normality (cont'd)

For the numerator:

$$\begin{aligned}\sqrt{n}\, l_n'(\theta_0) &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} l_i'(\theta_0) - 0\right)\\ &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} l_i'(\theta_0) - \mathbb{E}_{\theta_0} l_1'(\theta_0)\right)\\ &\to \mathcal{N}\left(0, \mathrm{Var}_{\theta_0}(l_1'(\theta_0)) = I(\theta_0)\right), \qquad \text{by CLT.}\end{aligned}$$

For the denominator:

- For all θ, $l_n''(\theta) \xrightarrow{P} \mathbb{E}_{\theta_0} l_1''(\theta)$ (by WLLN);
- Since $\theta_n^* \in [\theta_0, \hat{\theta}_n]$ and $\hat{\theta}_n \xrightarrow{P} \theta_0$ (by consistency), we have $\theta_n^* \xrightarrow{P} \theta_0$;
- Therefore $l_n''(\theta_n^*) \xrightarrow{P} \mathbb{E}_{\theta_0} l_1''(\theta_0) = -I(\theta_0)$.