Utilisation du pooling pour les tests RT-qPCR

Vincent Brault
sur un travail réalisé en collaboration avec Bastien Mallein et Jean-François Rupprecht

Mercredi 23 juin

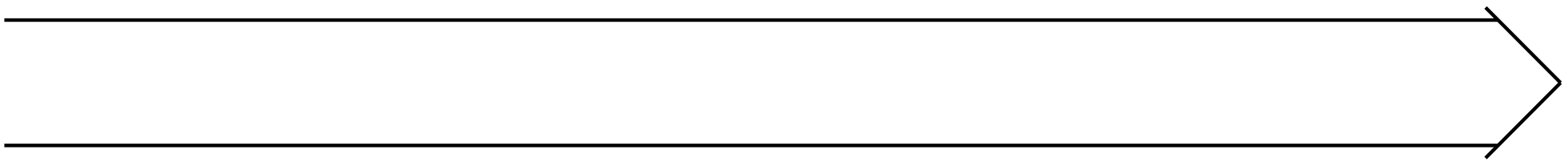
LABORATOIRE JEAN KUNTZMANN
MATHÉMATIQUES APPLIQUÉES - INFORMATIQUE

GROUP POOL

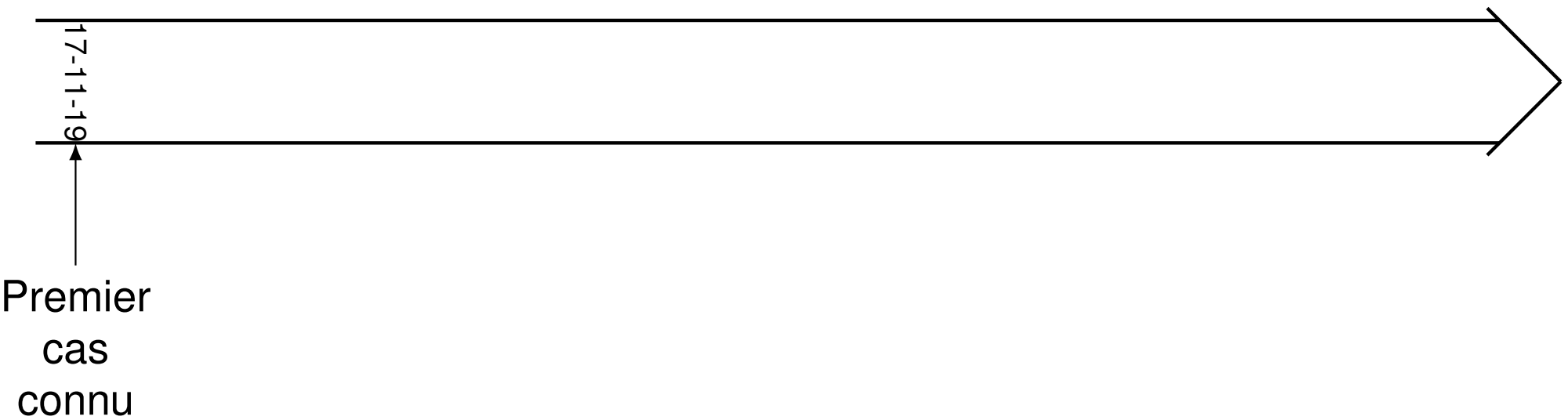
UGA Université Grenoble Alpes

Plan

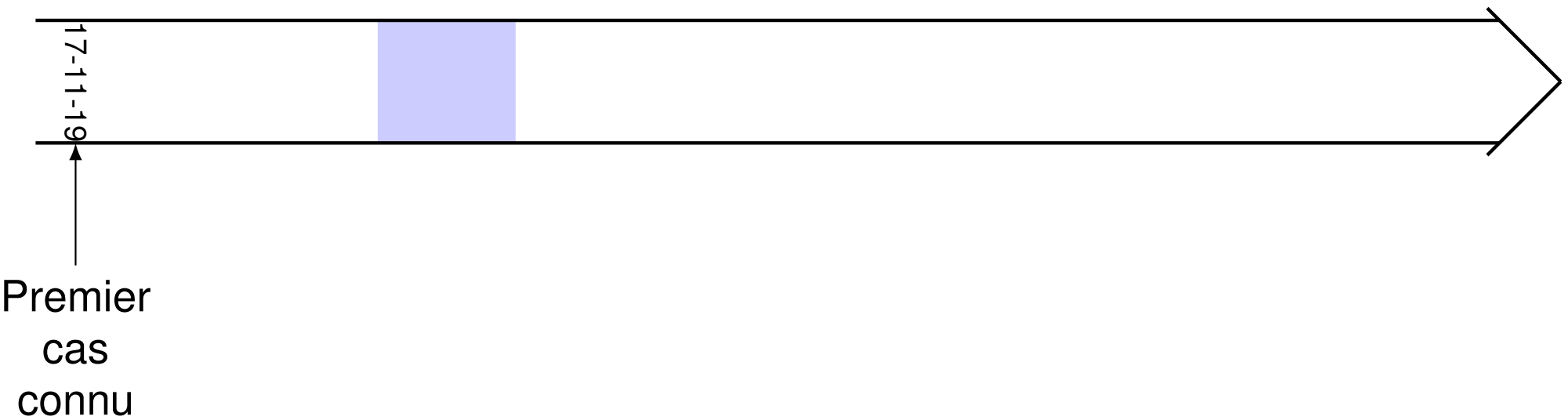
Chronologie sur les tests



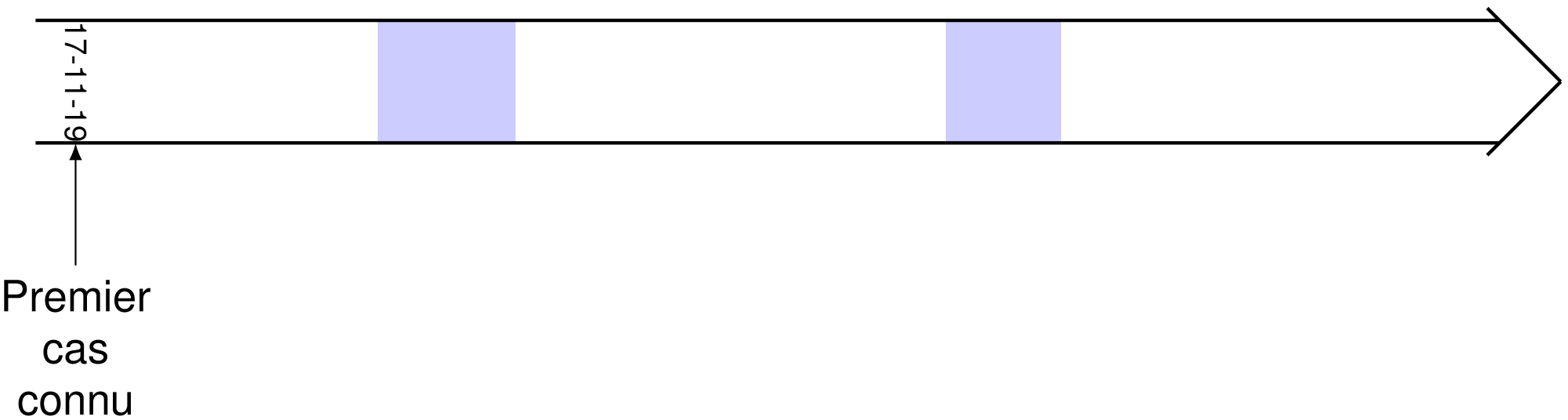
Chronologie sur les tests



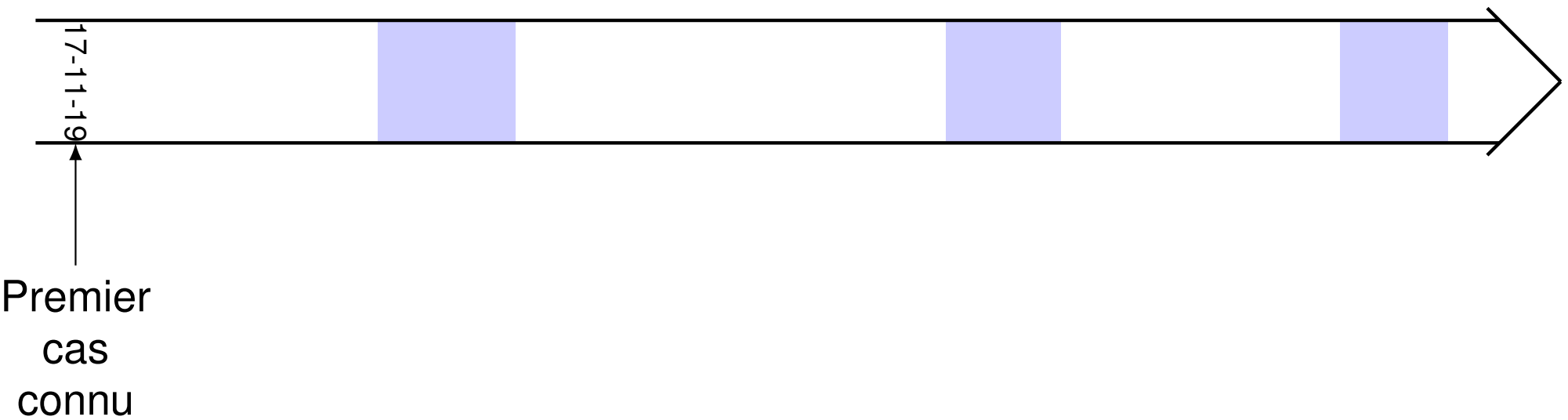
Chronologie sur les tests



Chronologie sur les tests



Chronologie sur les tests

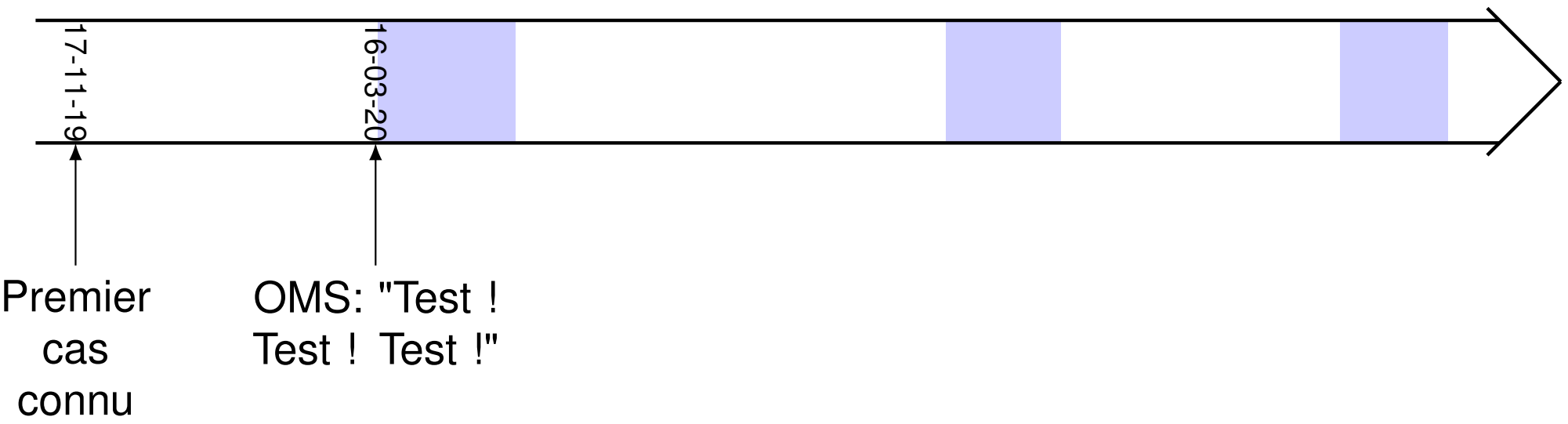


Chronologie sur les tests

Figure: Dr. Tedros Adhanom Ghebreyesus, Directeur général de l'OMS

Conseil de l'OMS, 16 mars 2020 : « Test ! Test ! Test ! »

Chronologie sur les tests



Chronologie sur les tests

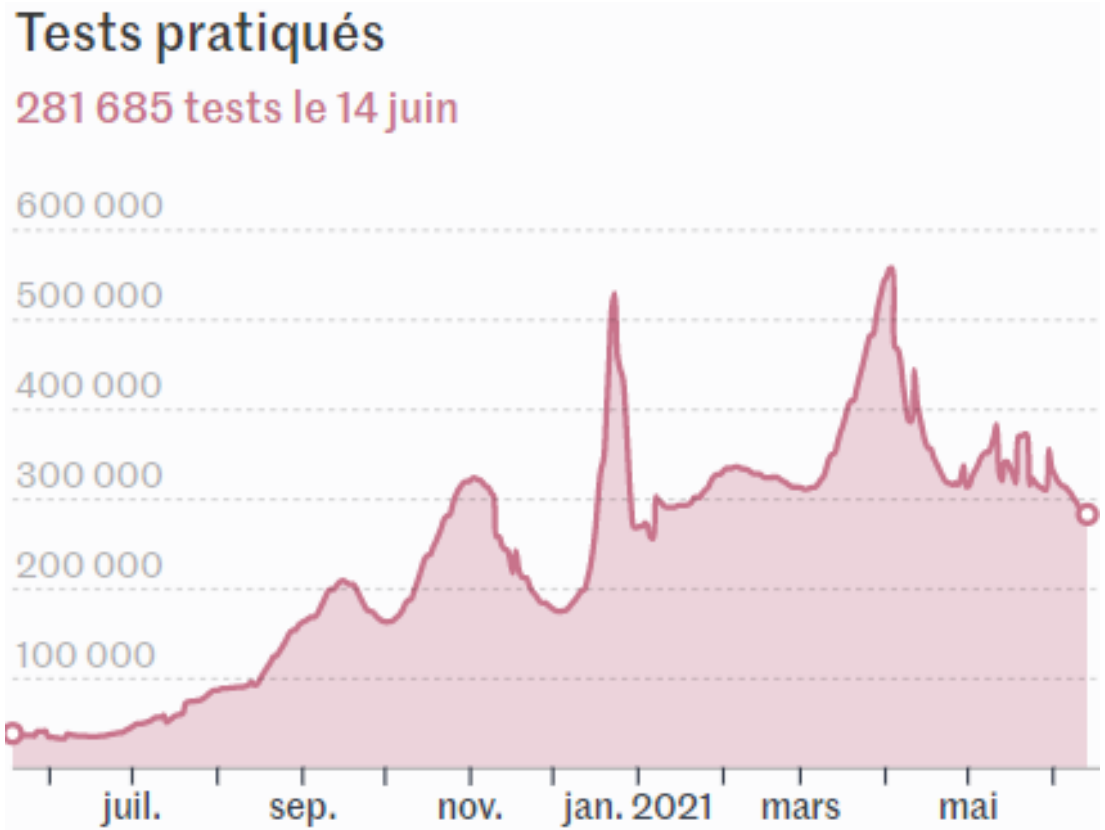
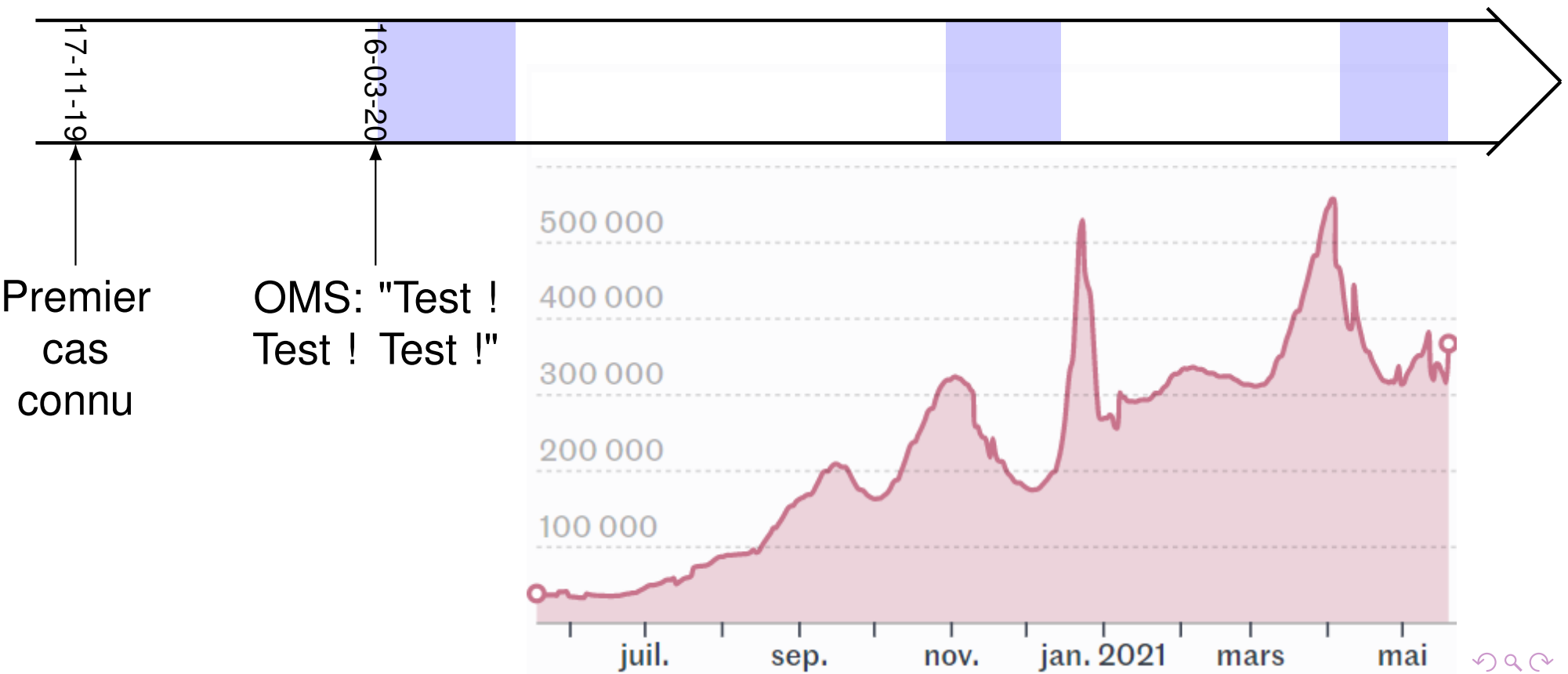


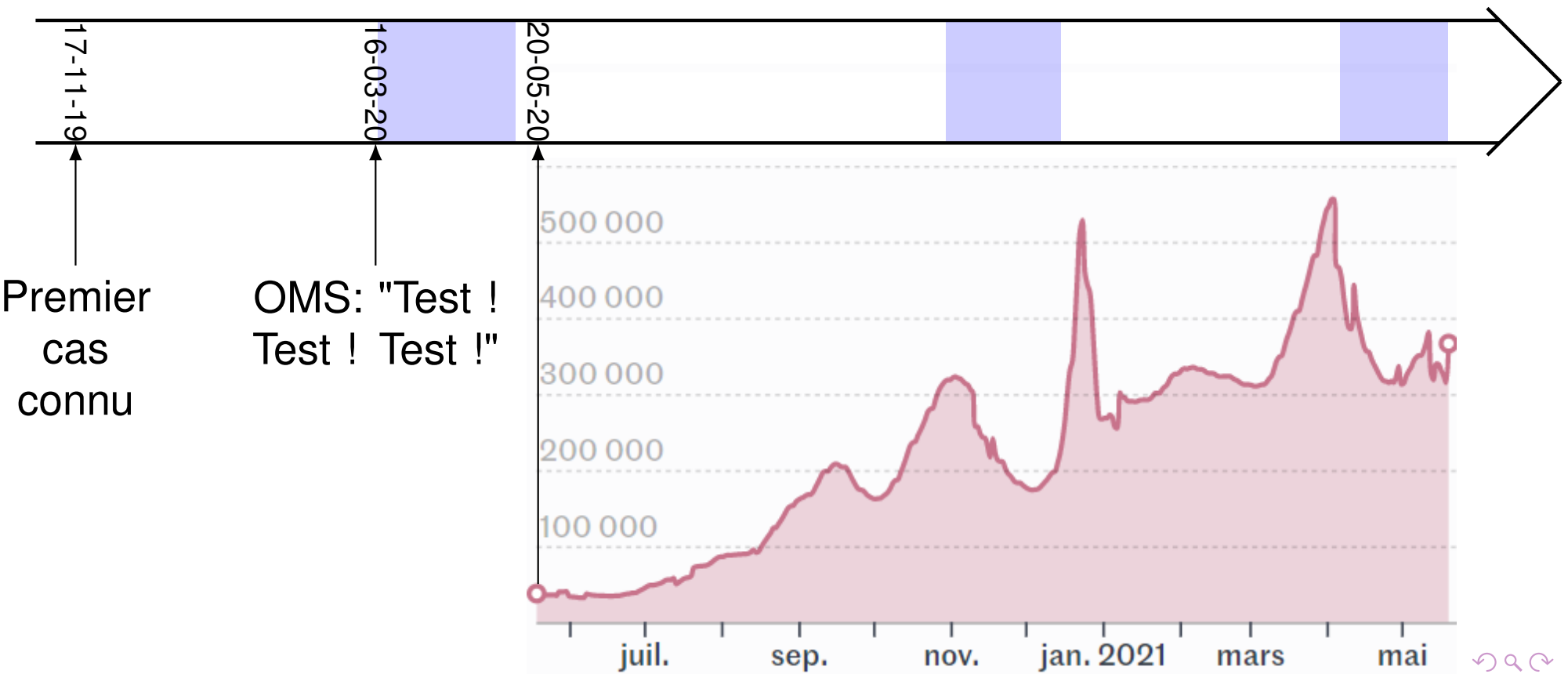
Figure: Evolution du nombre de tests en France entre le 20 mai 2020 et le 14 juin 2021

Source : Infographie du monde. https://www.lemonde.fr/les-decodeurs/article/2020/05/05/coronavirus-age-mortalite-departements-pays-suivez-l-evolution-de-l-epidemie-en-cartes-et-graphiques_6038751_4355770.html

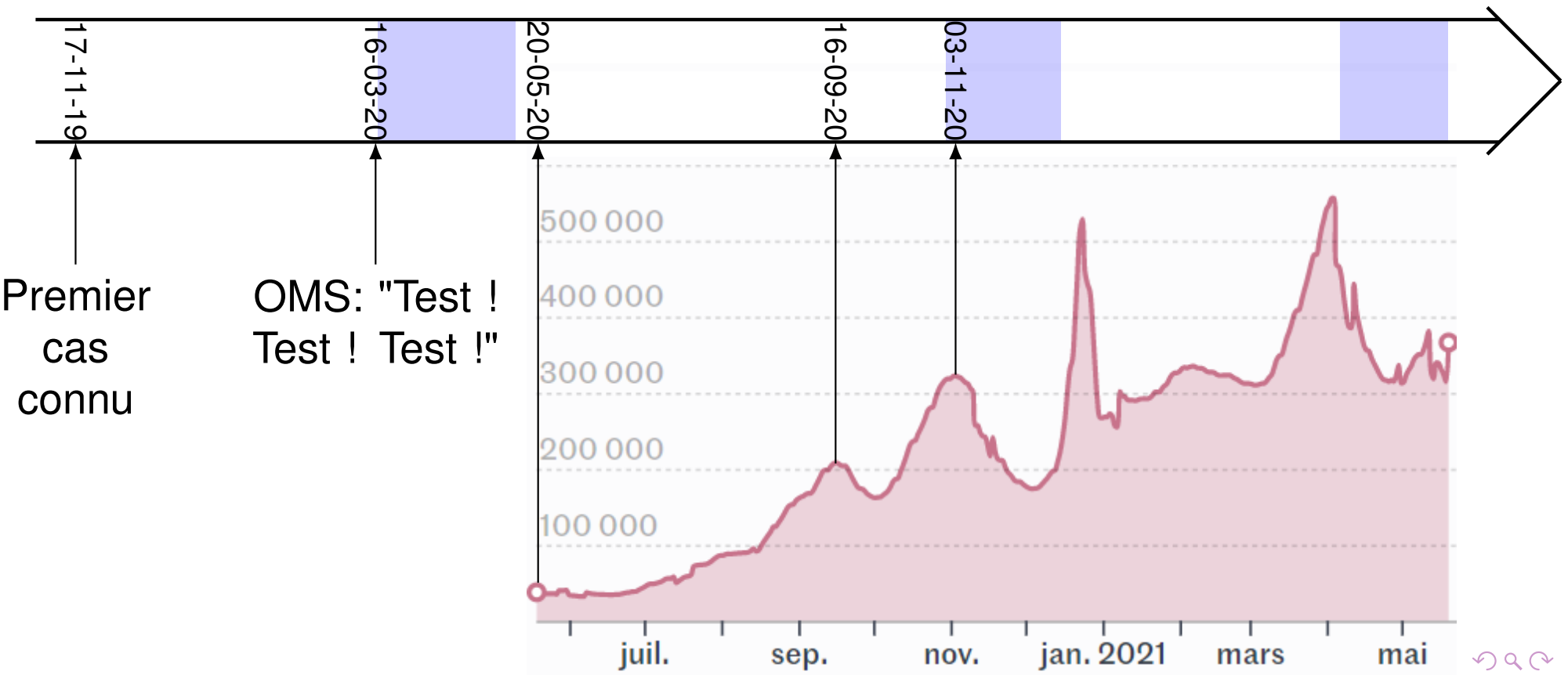
Chronologie sur les tests



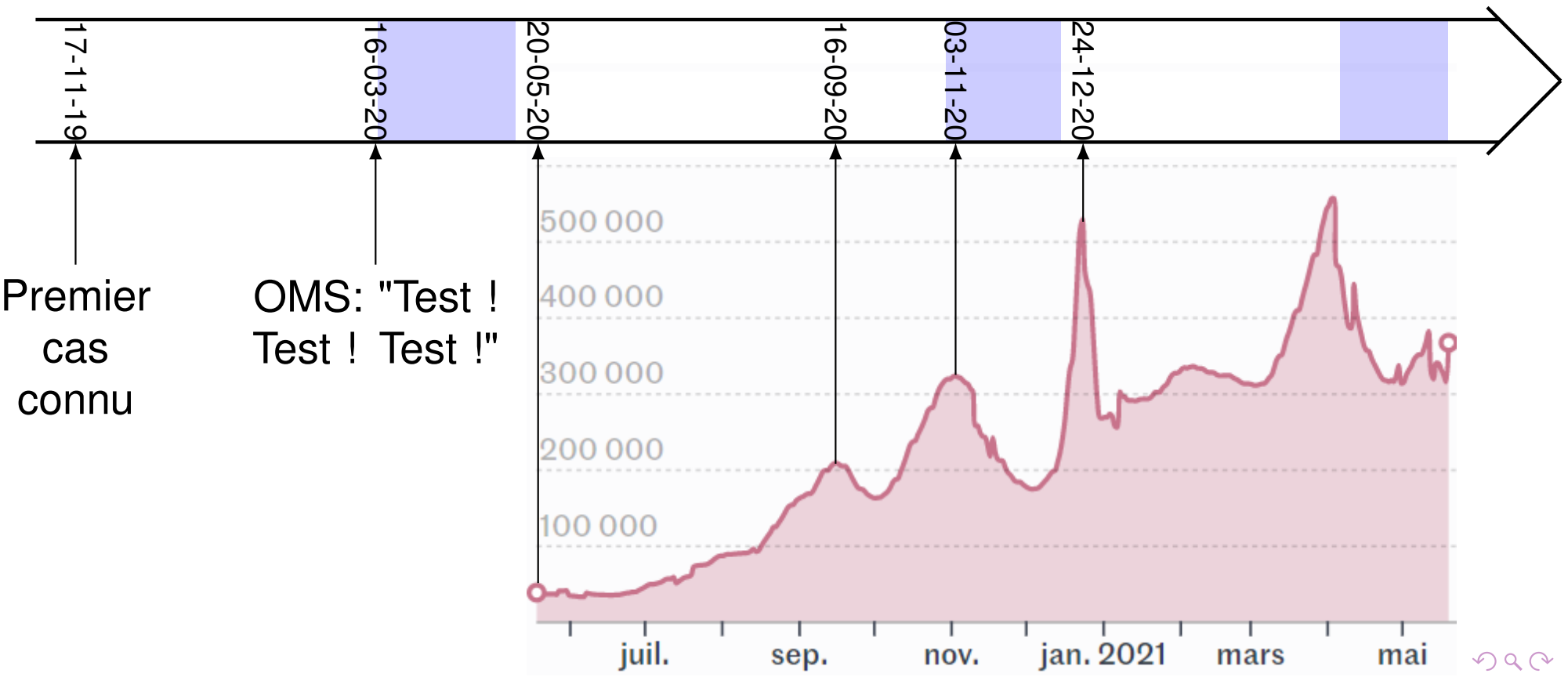
Chronologie sur les tests



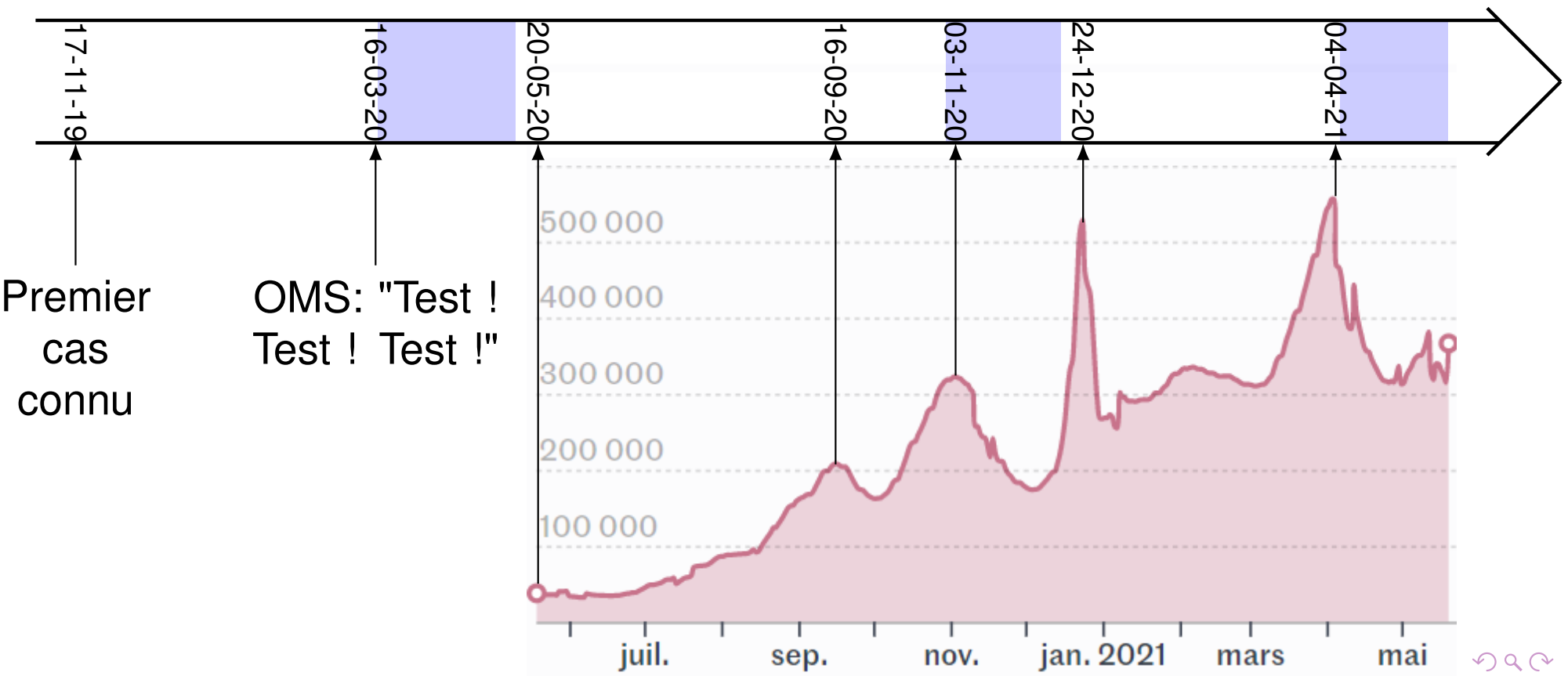
Chronologie sur les tests



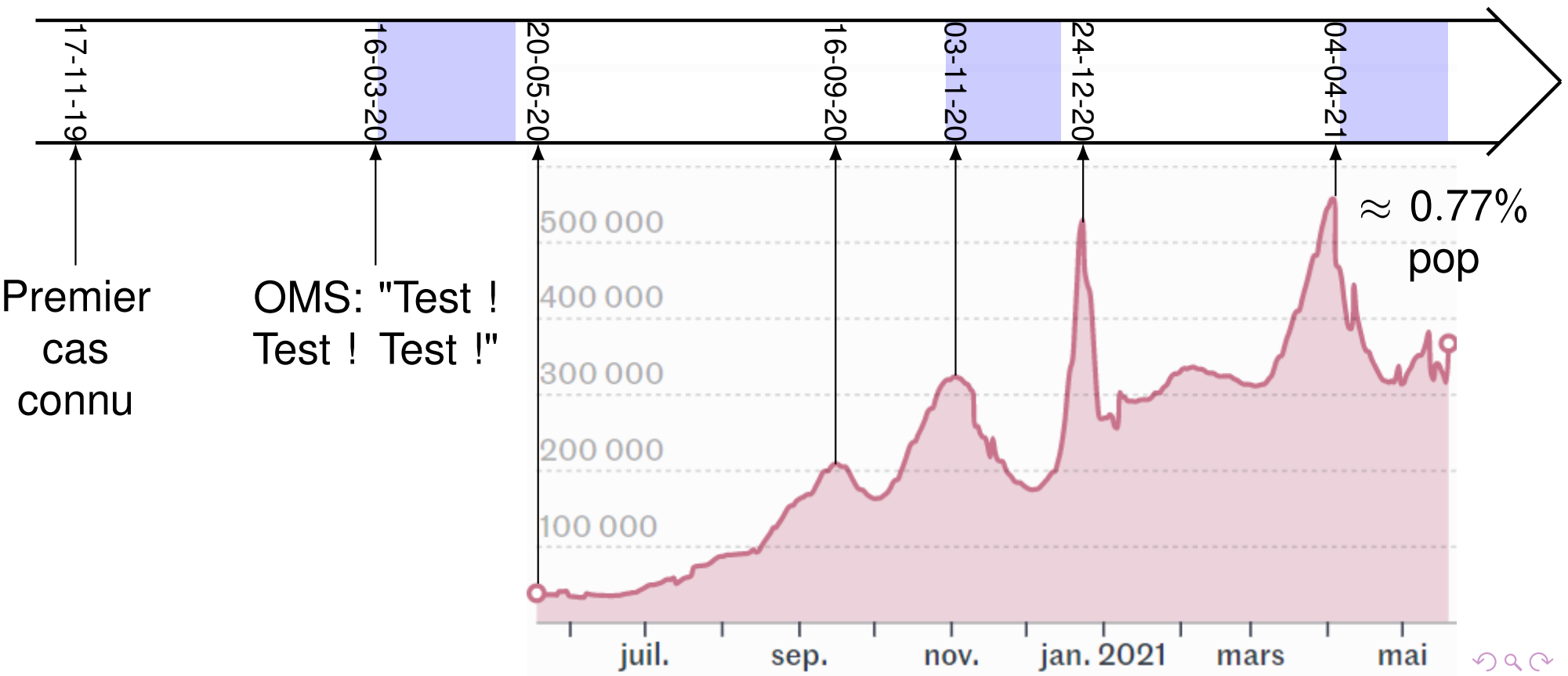
Chronologie sur les tests



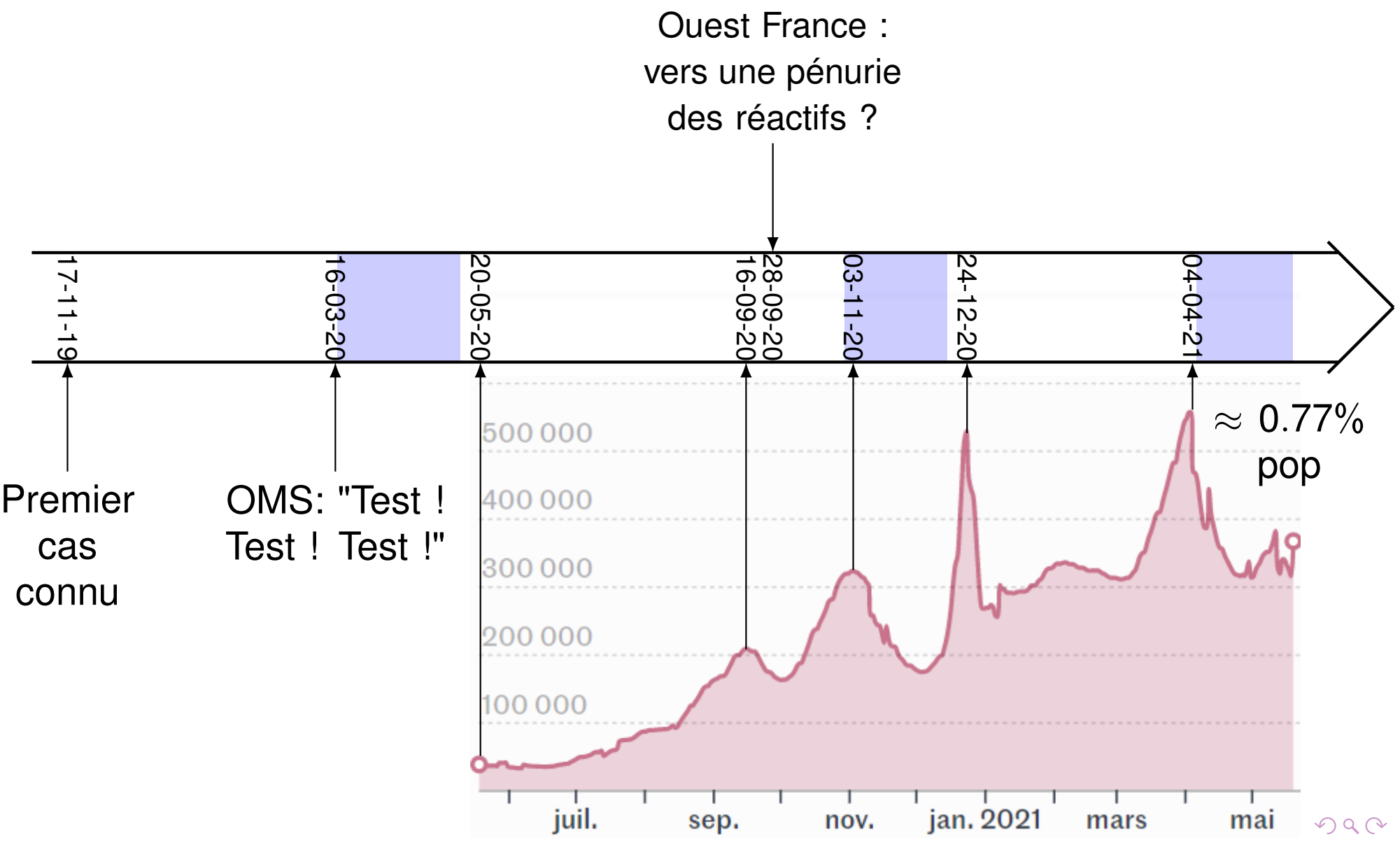
Chronologie sur les tests



Chronologie sur les tests



Chronologie sur les tests



Pourquoi tester ?

1. Estimer la prévalence[1] de COVID-19 dans une population.

[1] le nombre de cas dans une population.

Pourquoi tester ?

1. Estimer la prévalence[1] de COVID-19 dans une population.
2. Détecter rapidement la contamination d'une communauté fermée (prison ou maison de retraite par exemple)

[1] le nombre de cas dans une population.

Pourquoi tester ?

1. Estimer la prévalence[1] de COVID-19 dans une population.
2. Détecter rapidement la contamination d'une communauté fermée (prison ou maison de retraite par exemple)
3. Détecter rapidement et efficacement les individus contaminés.

[1] le nombre de cas dans une population.

Pourquoi tester ?

1. Estimer la prévalence[1] de COVID-19 dans une population.
2. Détecter rapidement la contamination d'une communauté fermée (prison ou maison de retraite par exemple)
3. Détecter rapidement et efficacement les individus contaminés.

Quelques problèmes :

1. Des cas asymptomatiques (combien ?)

[1] le nombre de cas dans une population.

Pourquoi tester ?

1. Estimer la prévalence[1] de COVID-19 dans une population.
2. Détecter rapidement la contamination d'une communauté fermée (prison ou maison de retraite par exemple)
3. Détecter rapidement et efficacement les individus contaminés.

Quelques problèmes :

1. Des cas asymptomatiques (combien ?)
2. Des individus contagieux avant les premiers symptômes

[1] le nombre de cas dans une population.

Pourquoi tester ?

1. Estimer la prévalence[1] de COVID-19 dans une population.
2. Détecter rapidement la contamination d'une communauté fermée (prison ou maison de retraite par exemple)
3. Détecter rapidement et efficacement les individus contaminés.

Quelques problèmes :

1. Des cas asymptomatiques (combien ?)
2. Des individus contagieux avant les premiers symptômes
3. Découragement de l'utilisation des tests (prix, pénurie, inconfort)

[1] le nombre de cas dans une population.

Pourquoi tester ?

1. Estimer la prévalence[1] de COVID-19 dans une population.
2. Détecter rapidement la contamination d'une communauté fermée (prison ou maison de retraite par exemple)
3. Détecter rapidement et efficacement les individus contaminés.

Quelques problèmes :

1. Des cas asymptomatiques (combien ?)
2. Des individus contagieux avant les premiers symptômes
3. Découragement de l'utilisation des tests (prix, pénurie, inconfort)
4. Une prévalence faible dans la population (moins de 5%)

[1] le nombre de cas dans une population.

Pourquoi tester ?

1. Estimer la prévalence[1] de COVID-19 dans une population.
2. Détecter rapidement la contamination d'une communauté fermée (prison ou maison de retraite par exemple)
3. Détecter rapidement et efficacement les individus contaminés.

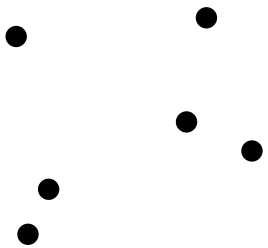
Quelques problèmes :

1. Des cas asymptomatiques (combien ?)
2. Des individus contagieux avant les premiers symptômes
3. Découragement de l'utilisation des tests (prix, pénurie, inconfort)
4. Une prévalence faible dans la population (moins de 5%)

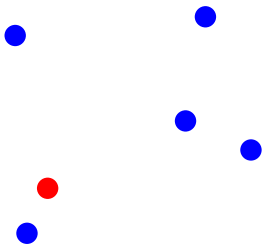
$\hookrightarrow$ utilisation du pooling

[1] le nombre de cas dans une population.

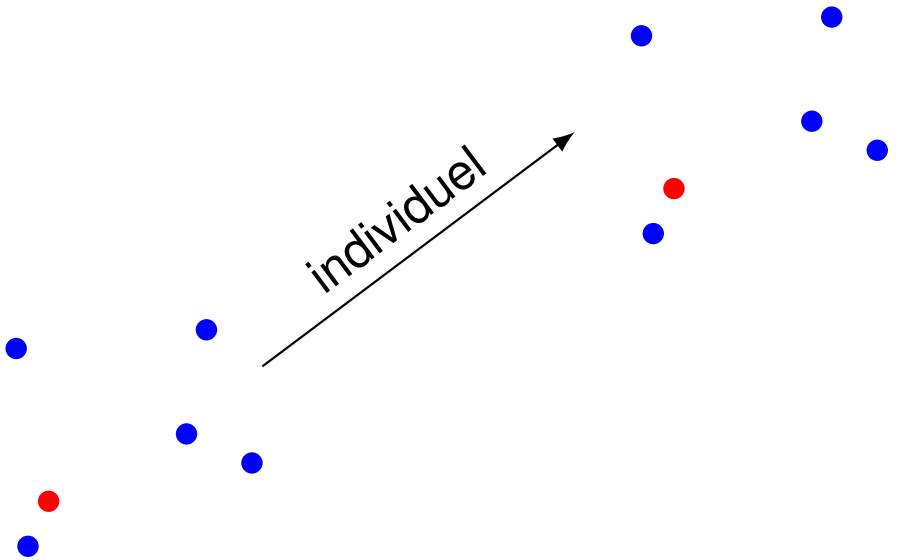
Le pooling ?



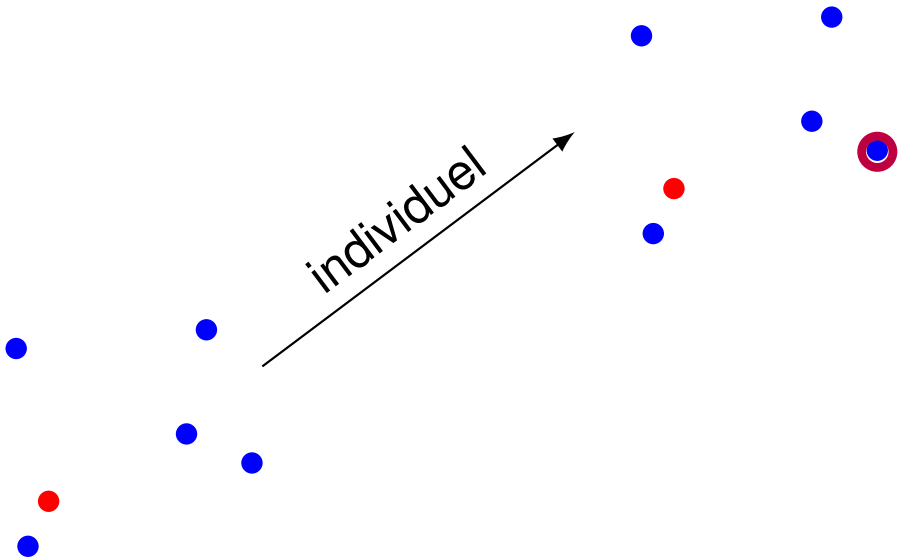
Le pooling ?



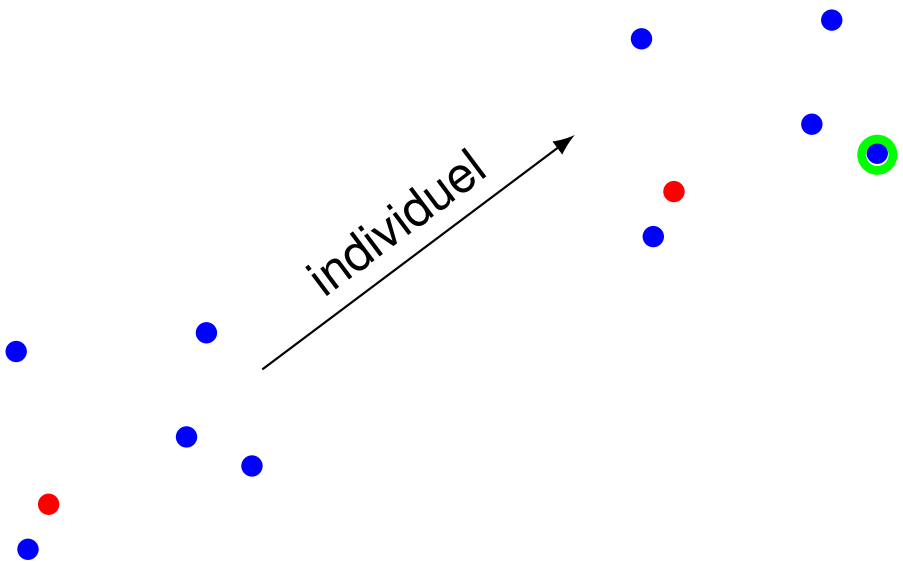
Le pooling ?



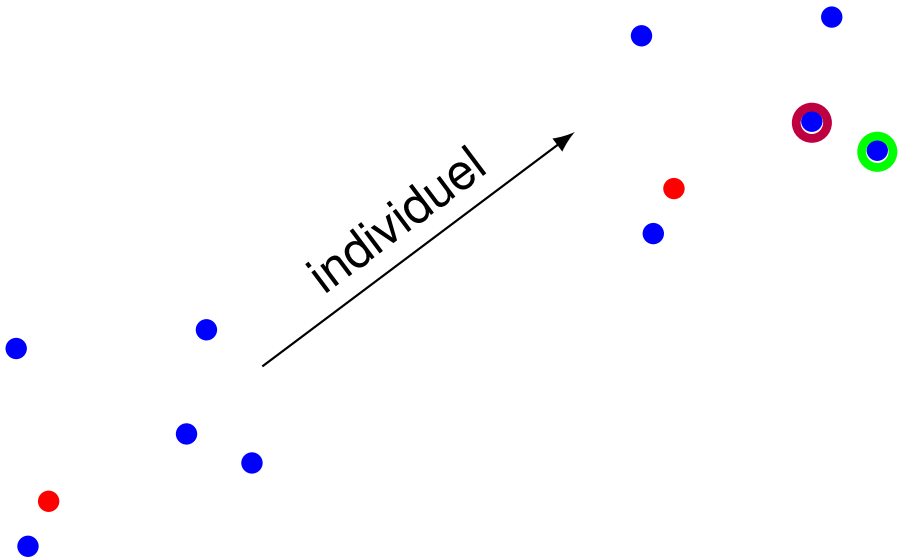
Le pooling ?



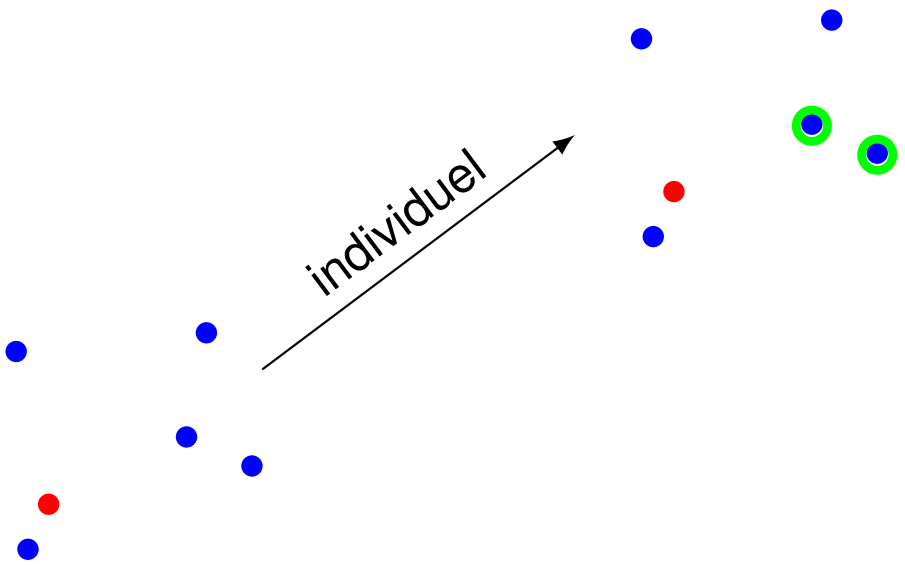
Le pooling ?



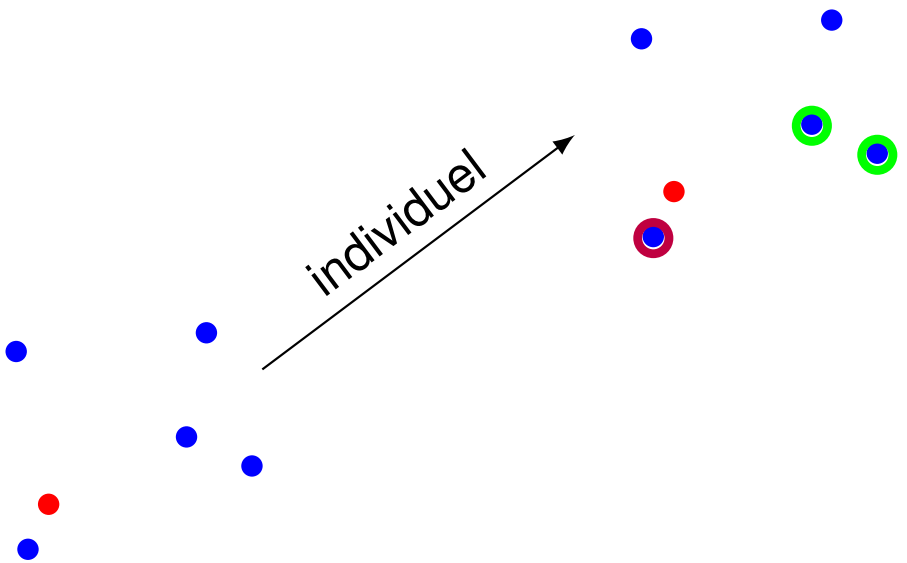
Le pooling ?



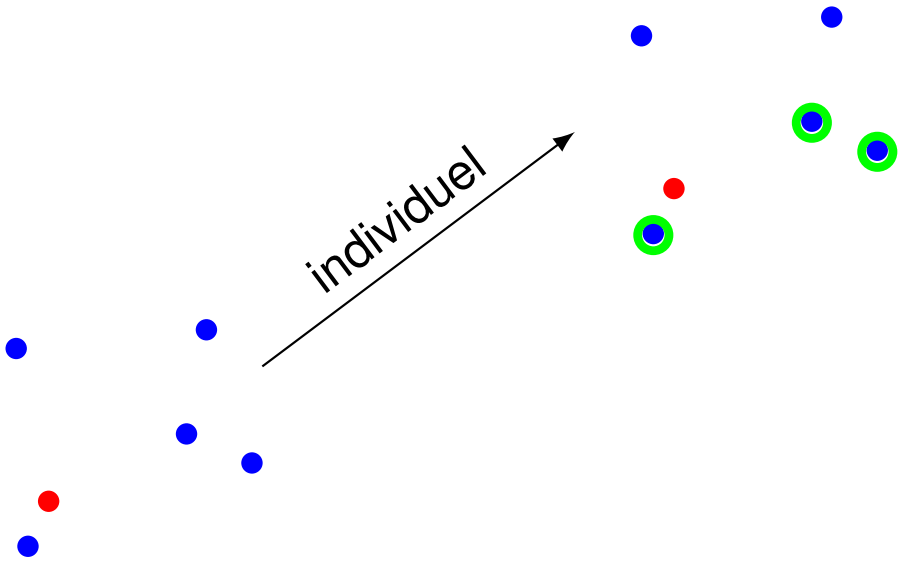
Le pooling ?



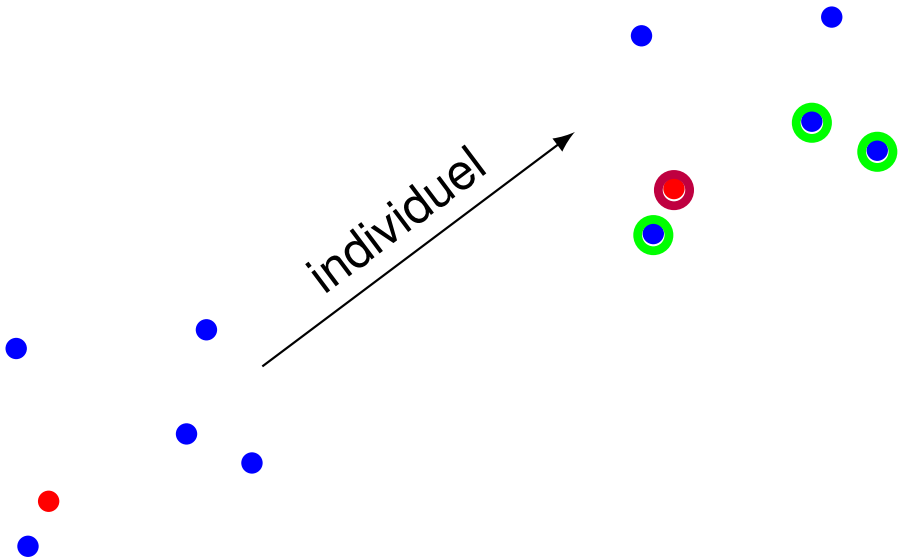
Le pooling ?



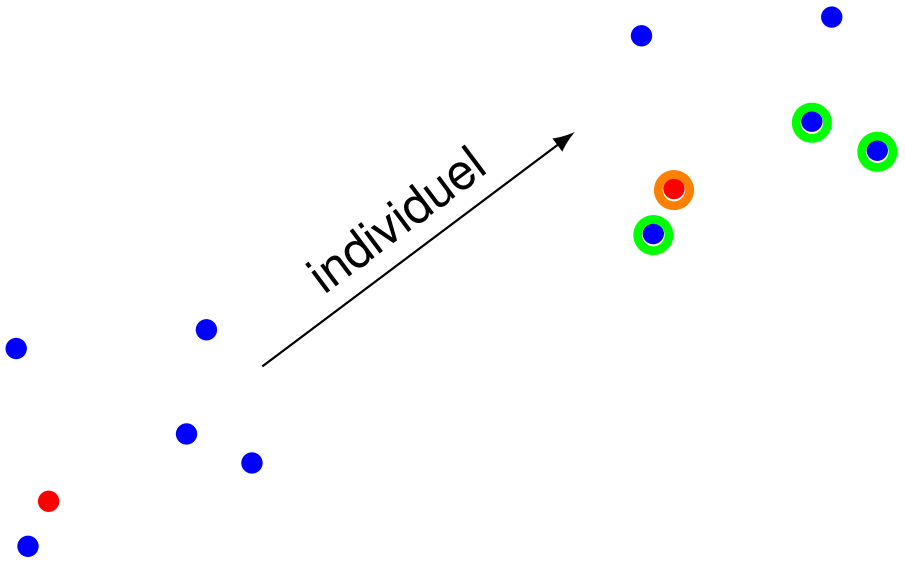
Le pooling ?



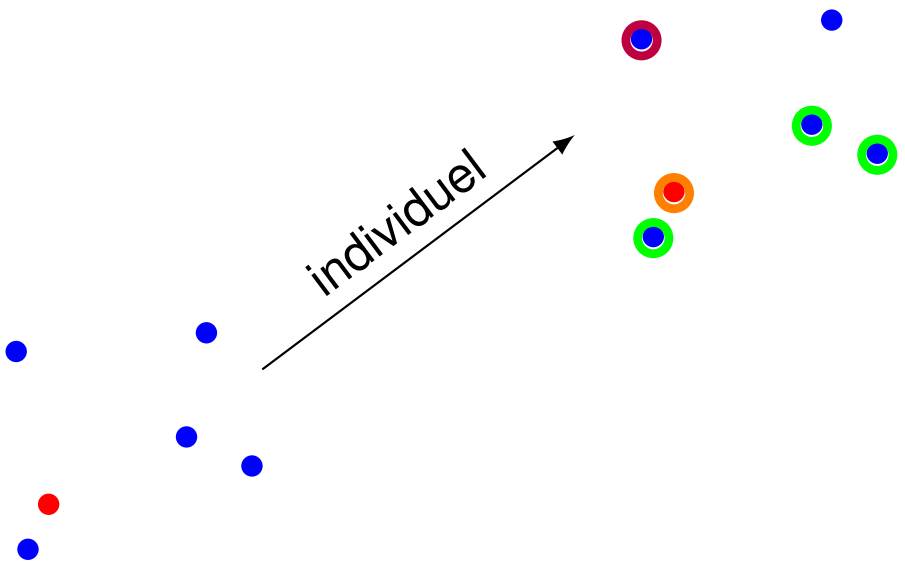
Le pooling ?



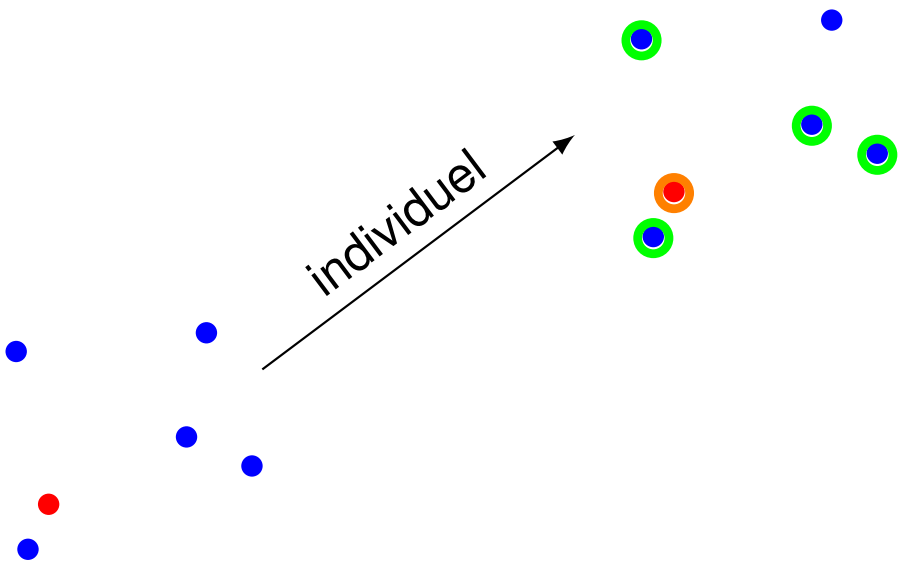
Le pooling ?



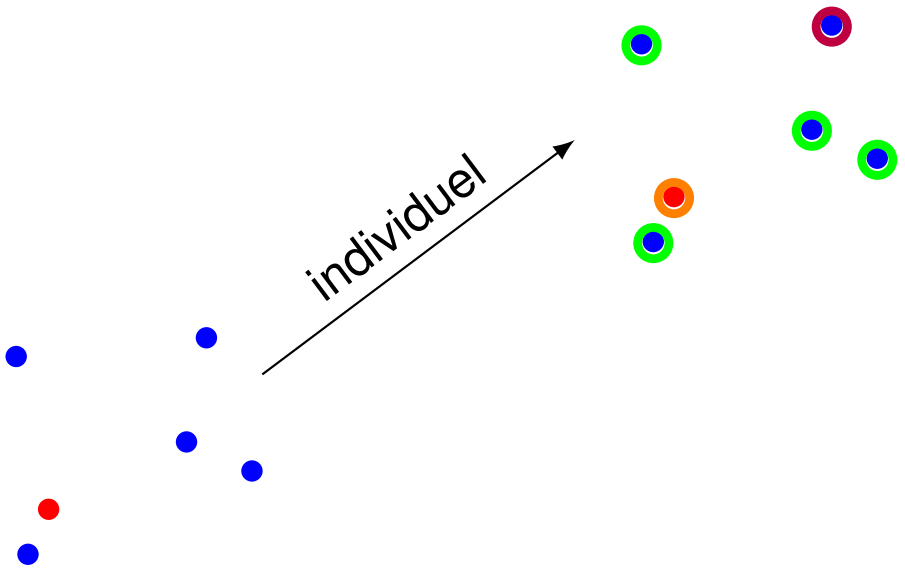
Le pooling ?



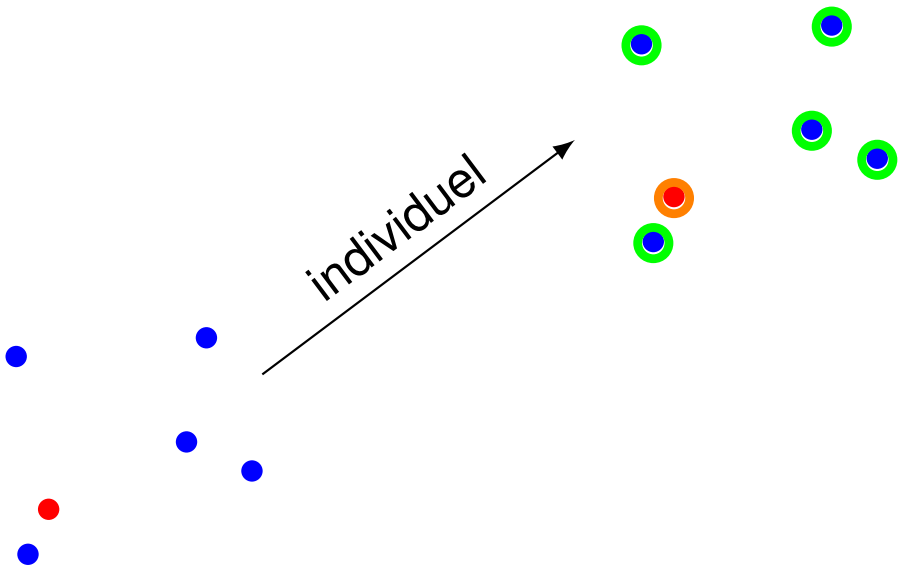
Le pooling ?



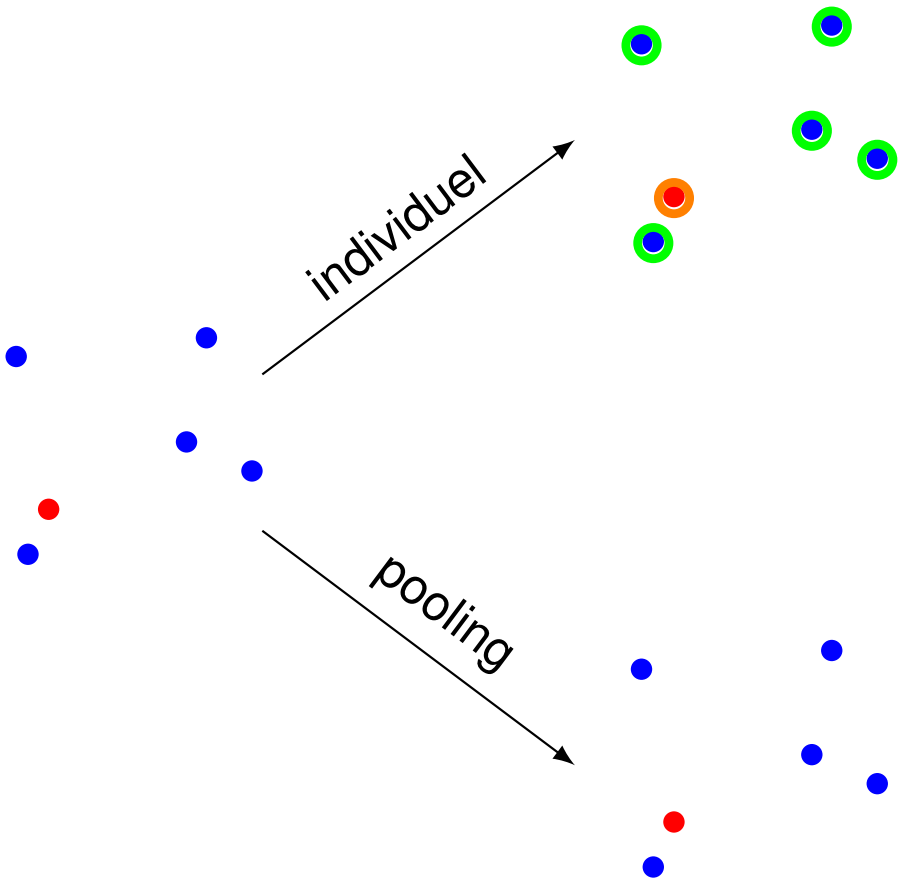
Le pooling ?



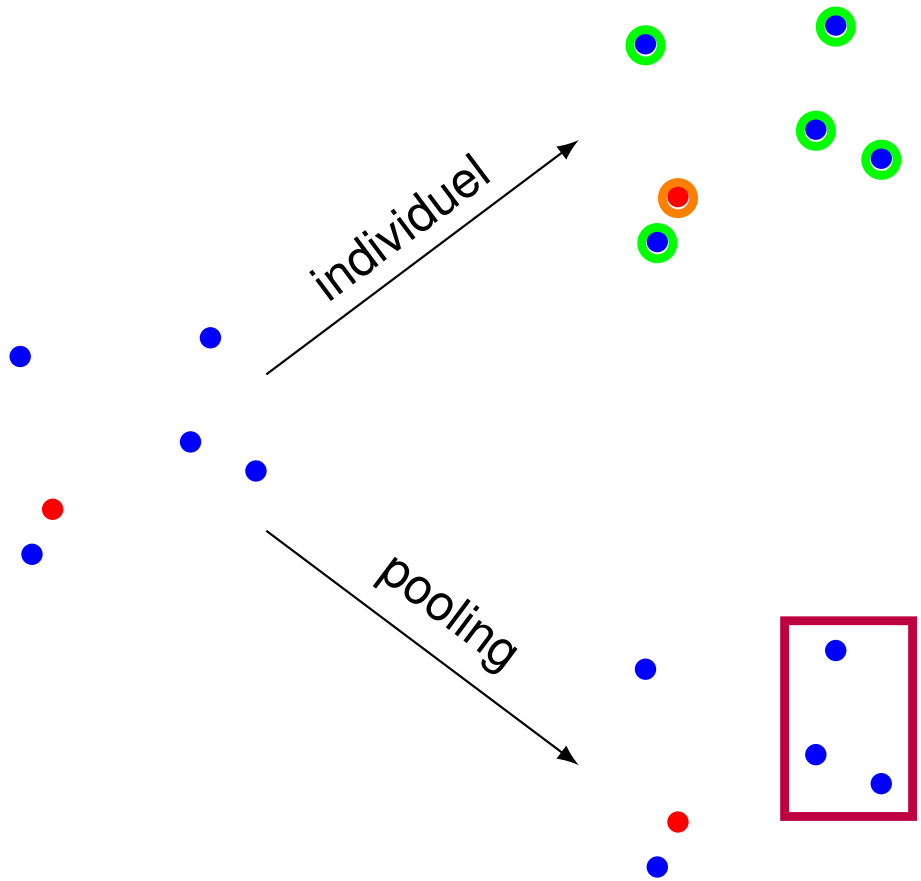
Le pooling ?



Le pooling ?

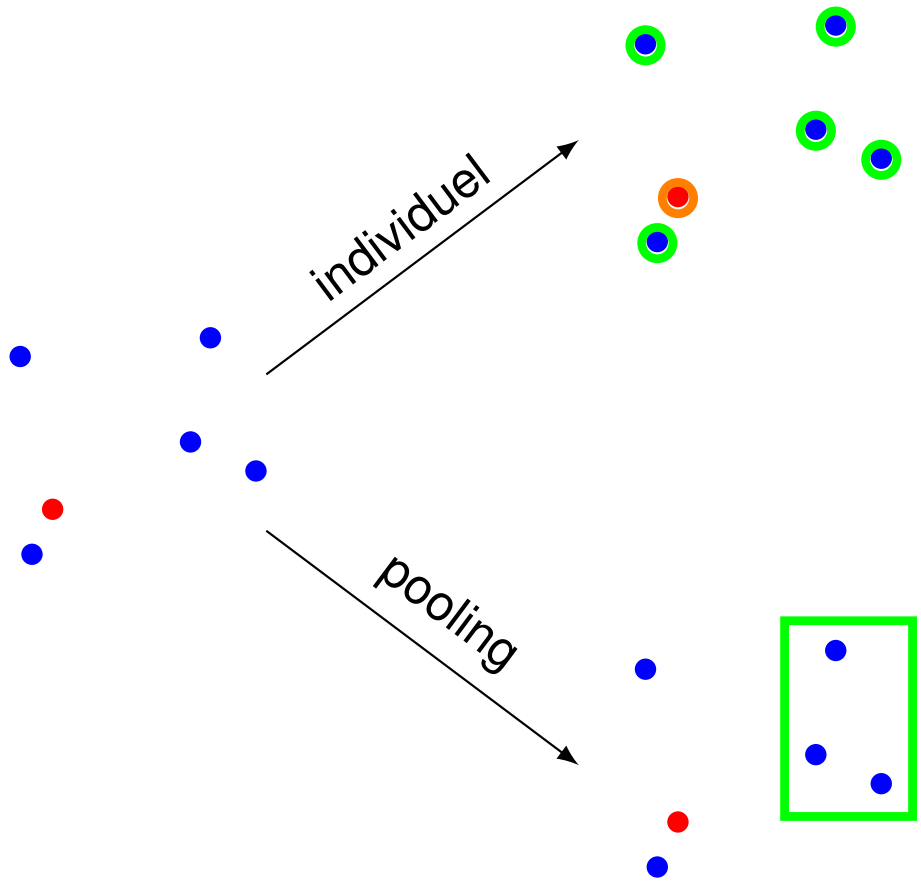


Le pooling ?



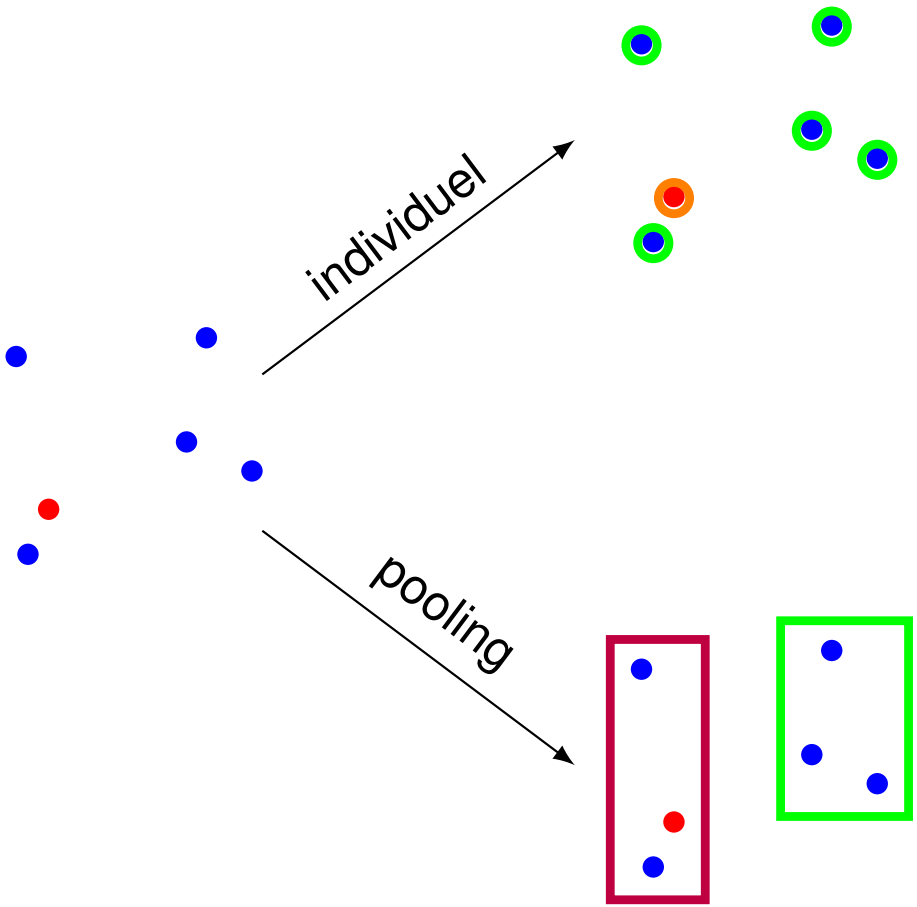
Mélange des échantillons puis un seul test.
Le groupe est négatif si et seulement si tous les membres sont négatifs.

Le pooling ?



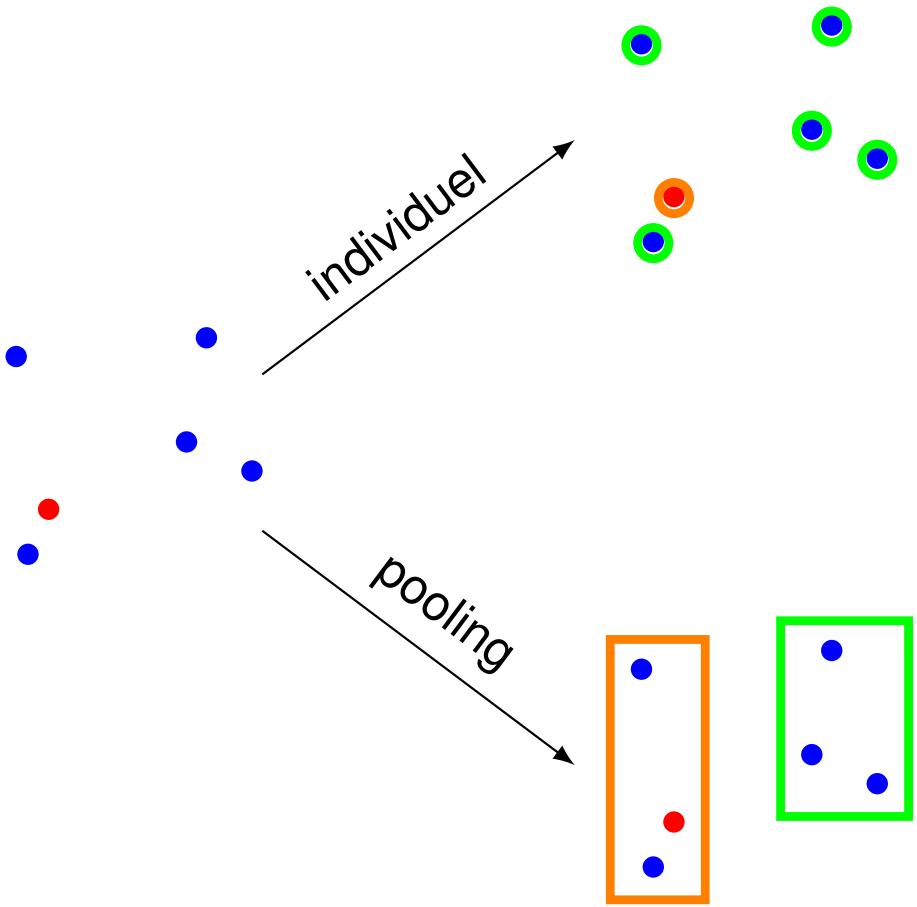
Mélange des échantillons puis un seul test.
Le groupe est négatif si et seulement si tous les membres sont négatifs.

Le pooling ?



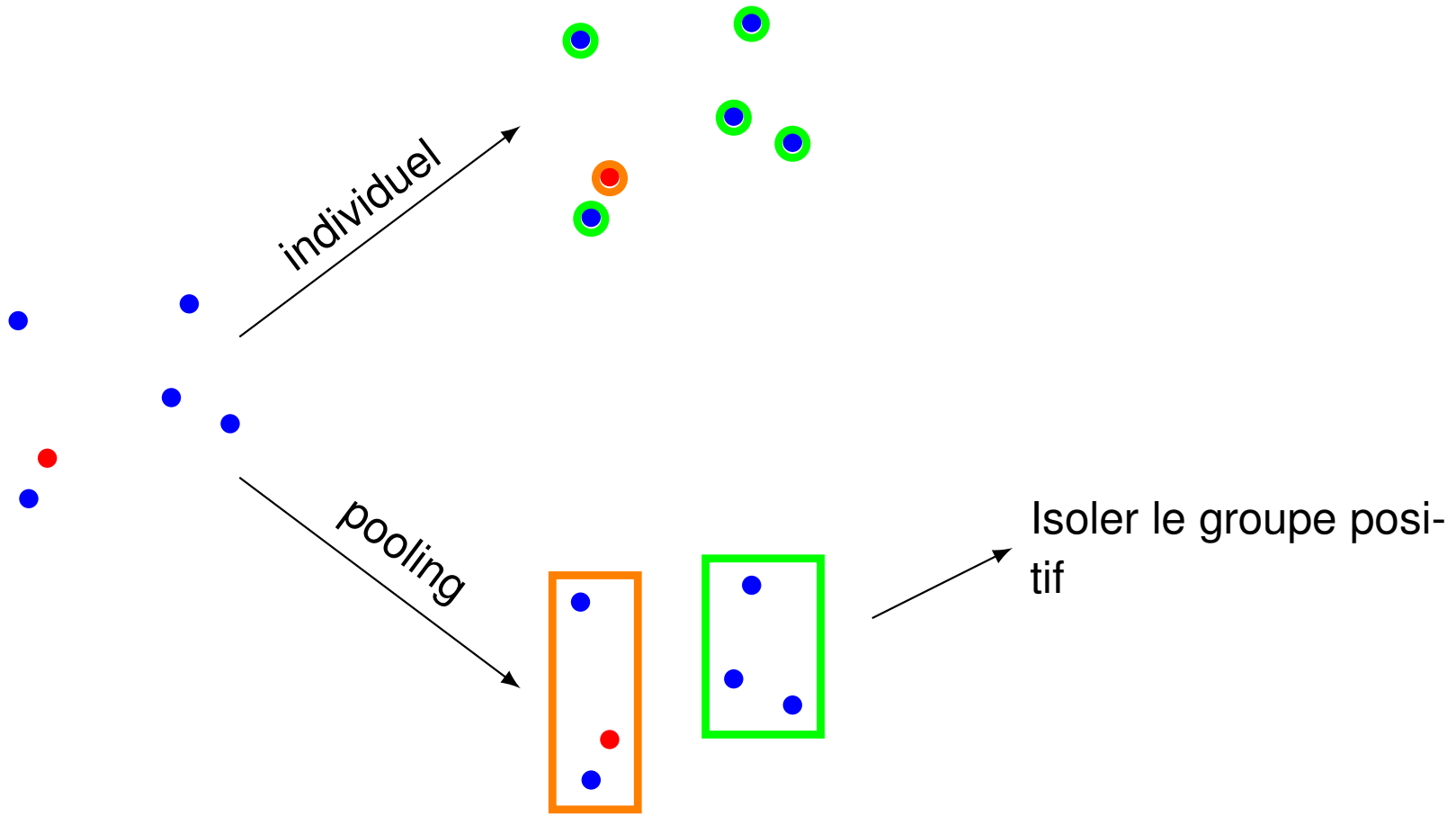
Mélange des échantillons puis un seul test.
Le groupe est négatif si et seulement si tous les membres sont négatifs.

Le pooling ?



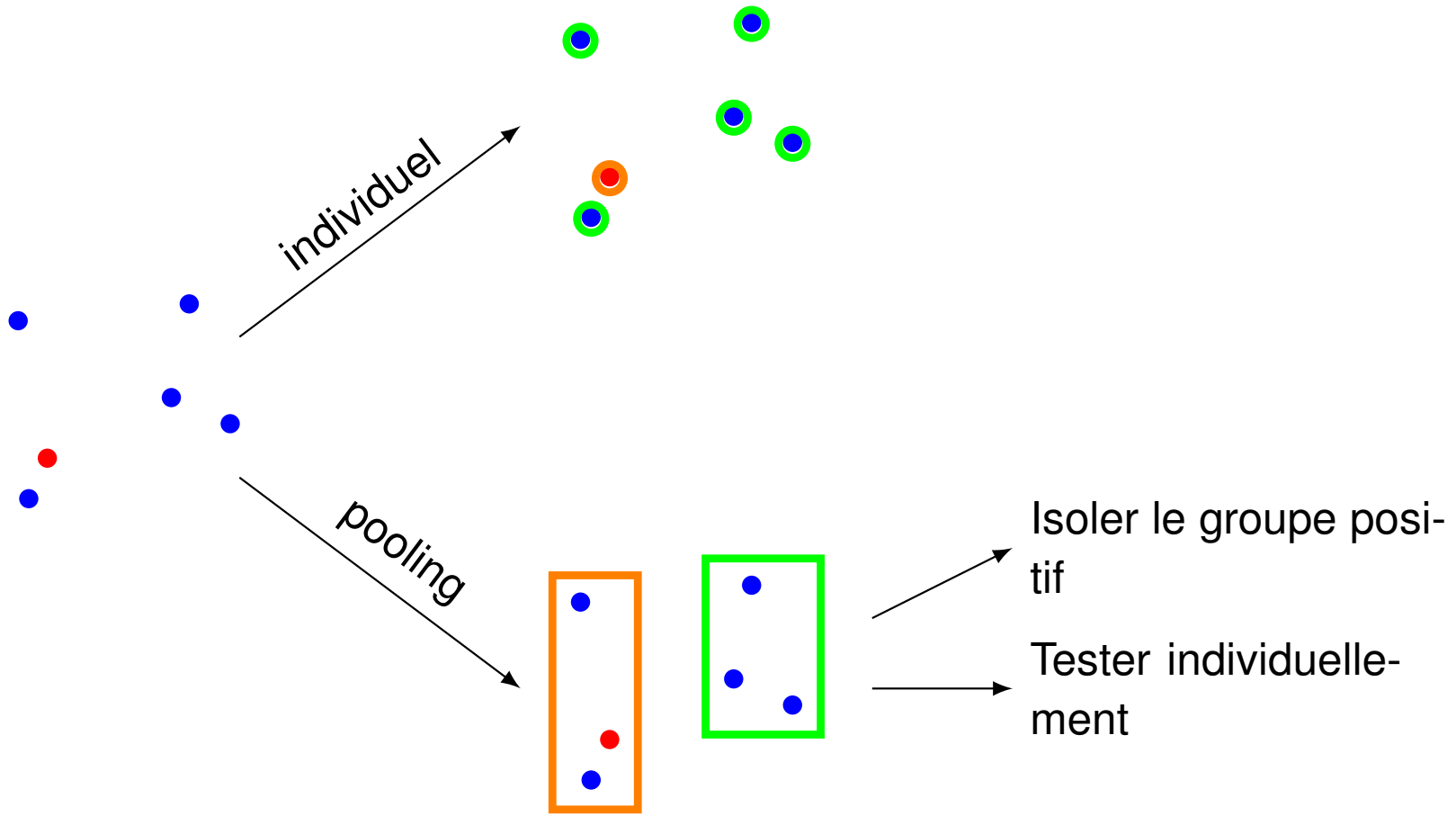
Mélange des échantillons puis un seul test.
Le groupe est négatif si et seulement si tous les membres sont négatifs.

Le pooling ?



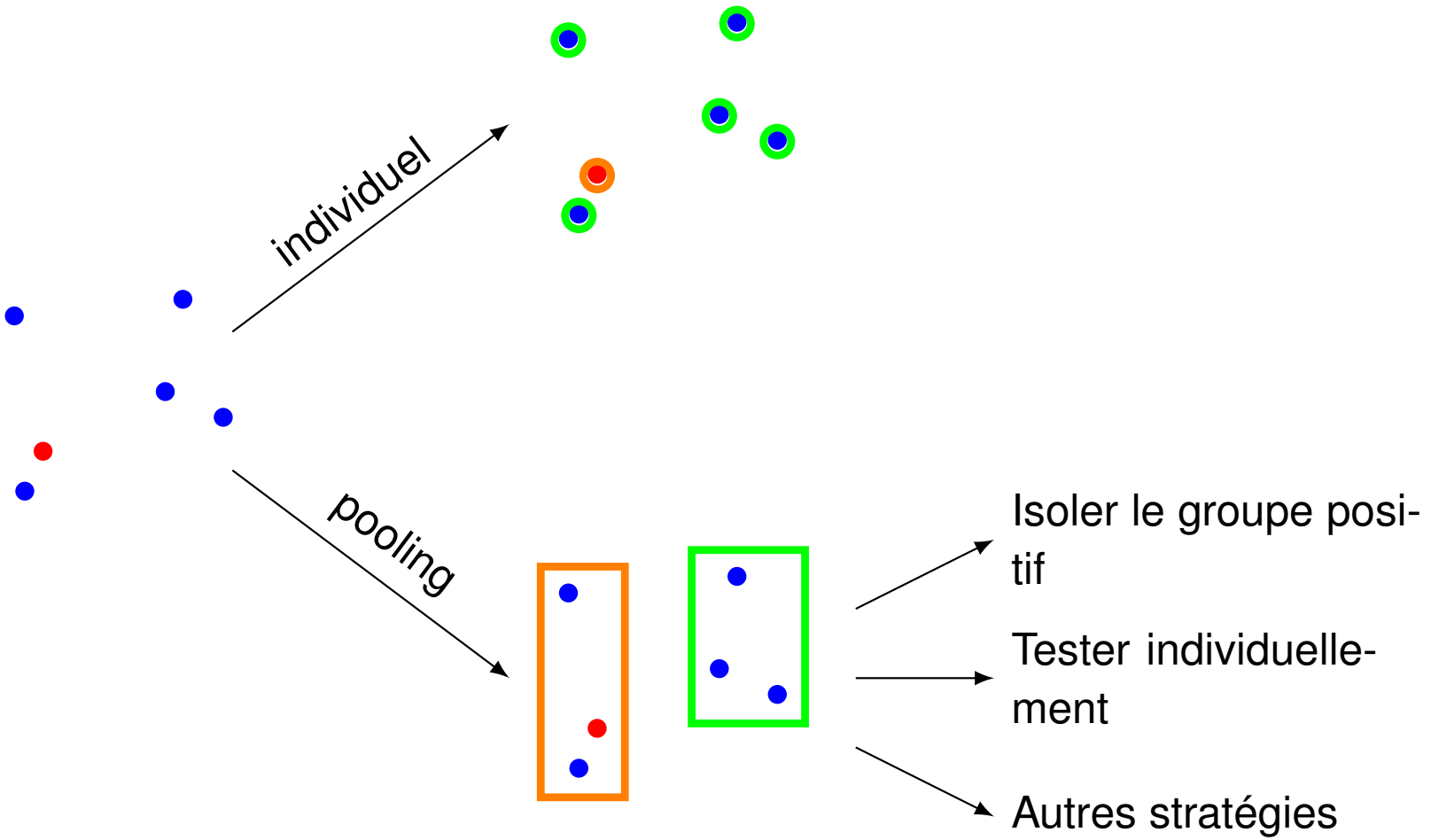
Mélange des échantillons puis un seul test.
Le groupe est négatif si et seulement si tous les membres sont négatifs.

Le pooling ?



Mélange des échantillons puis un seul test.
Le groupe est négatif si et seulement si tous les membres sont négatifs.

Le pooling ?



Mélange des échantillons puis un seul test.
Le groupe est négatif si et seulement si tous les membres sont négatifs.

Historique du pooling

1. Utilisé par Dorfman [1943] pour détecter la syphilis chez les soldats américains : groupes de taille $n = 128$ puis deuxième test individuel de toutes les personnes appartenant à un groupe contaminé.

Historique du pooling

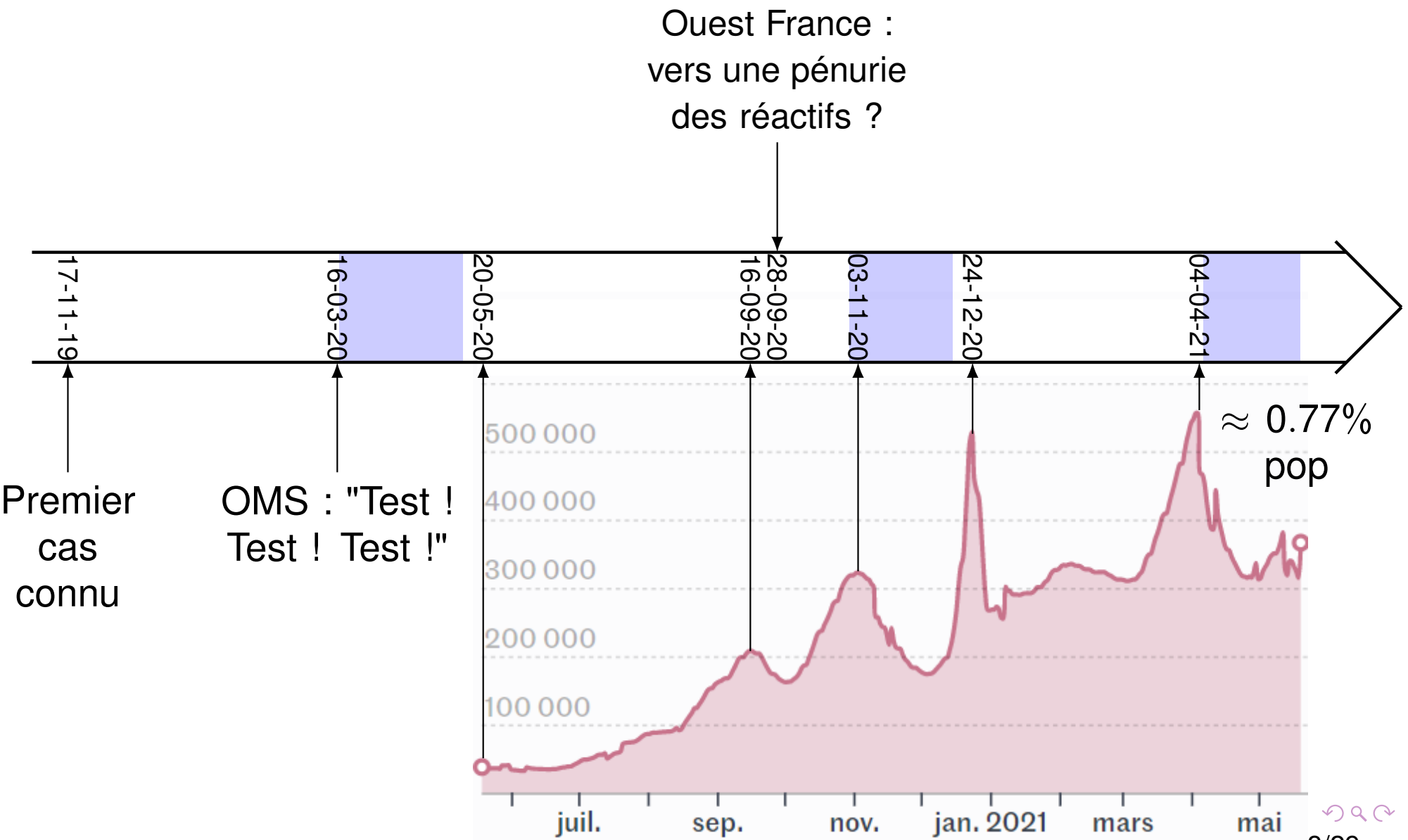
1. Utilisé par Dorfman [1943] pour détecter la syphilis chez les soldats américains : groupes de taille $n = 128$ puis deuxième test individuel de toutes les personnes appartenant à un groupe contaminé.
2. Utilisé en France pour détecter le VIH dans les produits du sang. Lié à l'affaire du sang contaminé ?

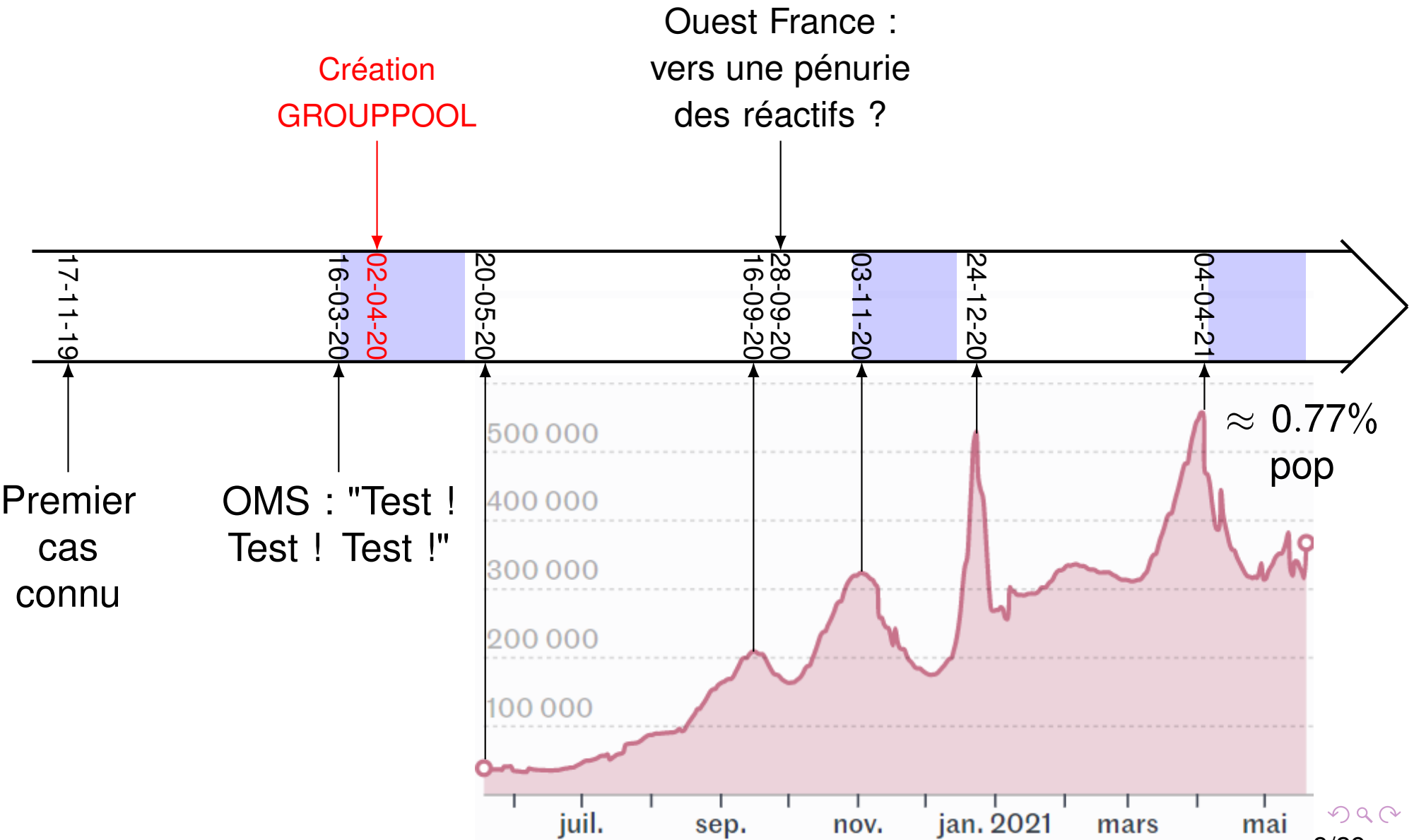
Historique du pooling

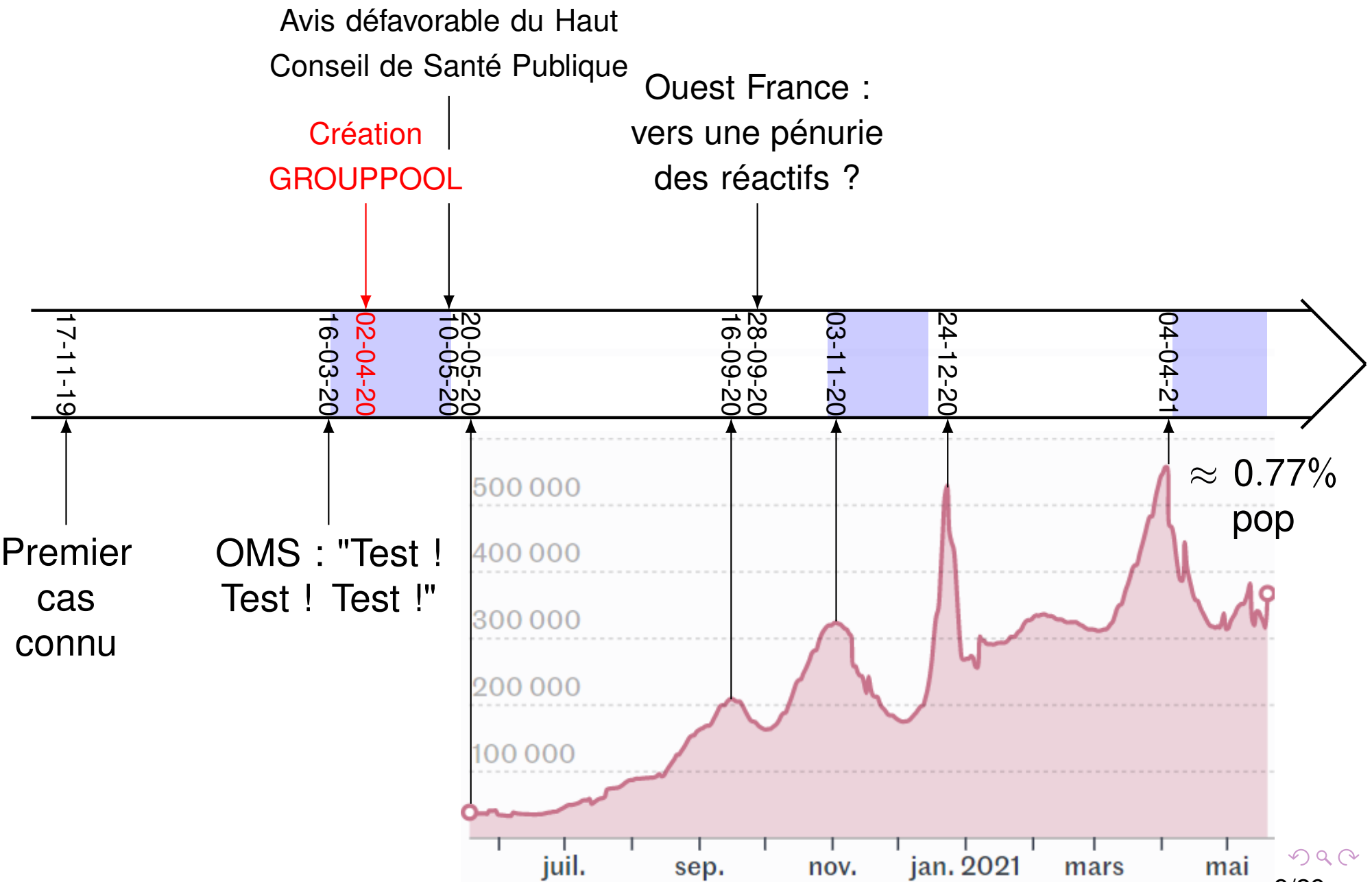
1. Utilisé par Dorfman [1943] pour détecter la syphilis chez les soldats américains : groupes de taille $n = 128$ puis deuxième test individuel de toutes les personnes appartenant à un groupe contaminé.
2. Utilisé en France pour détecter le VIH dans les produits du sang. Lié à l'affaire du sang contaminé ?
3. Envisagé pendant l'épidémie de SARS-COV-1 pour mettre en quarantaine régulièrement des groupes de patients suspects.

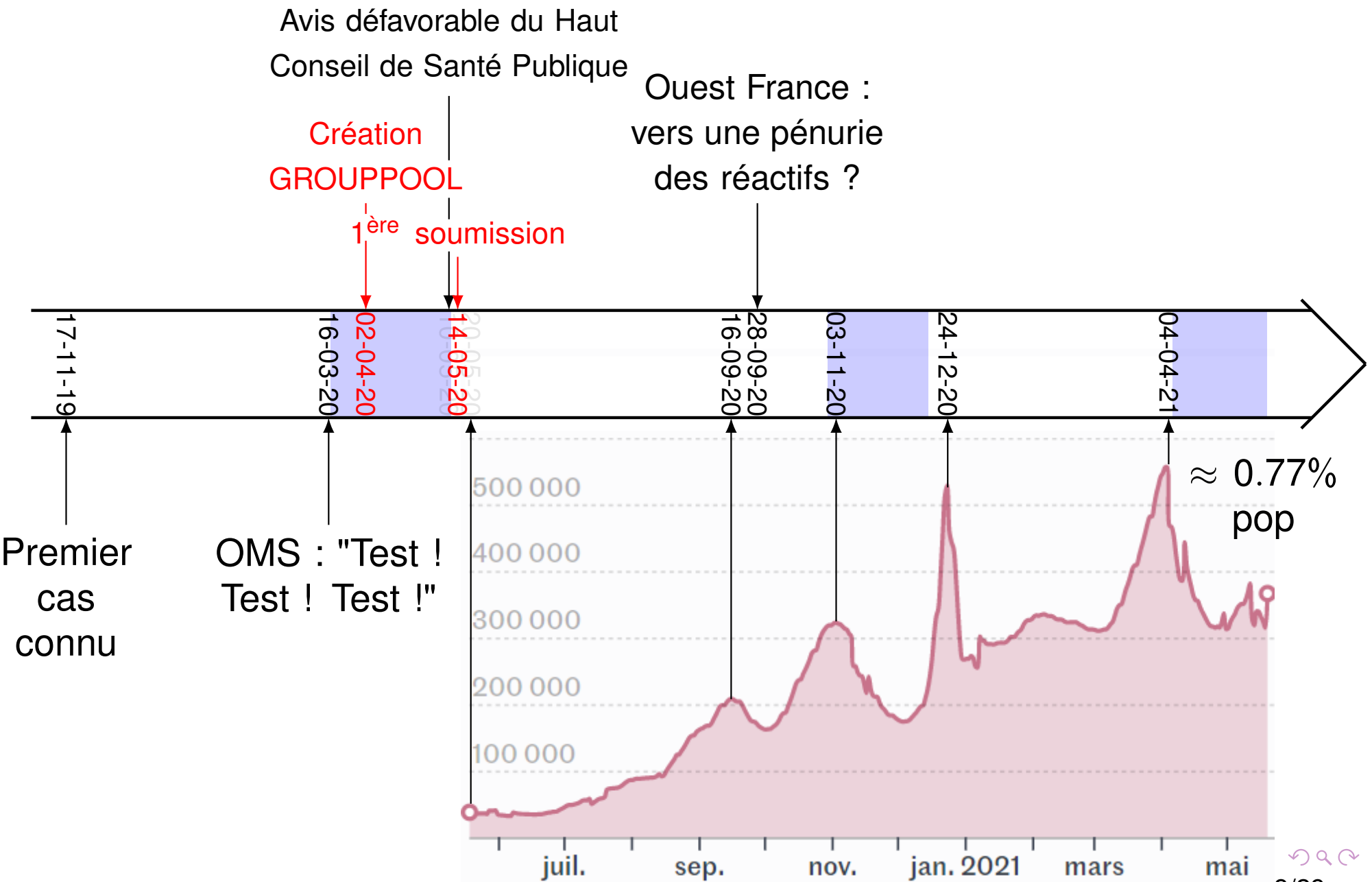
Historique du pooling

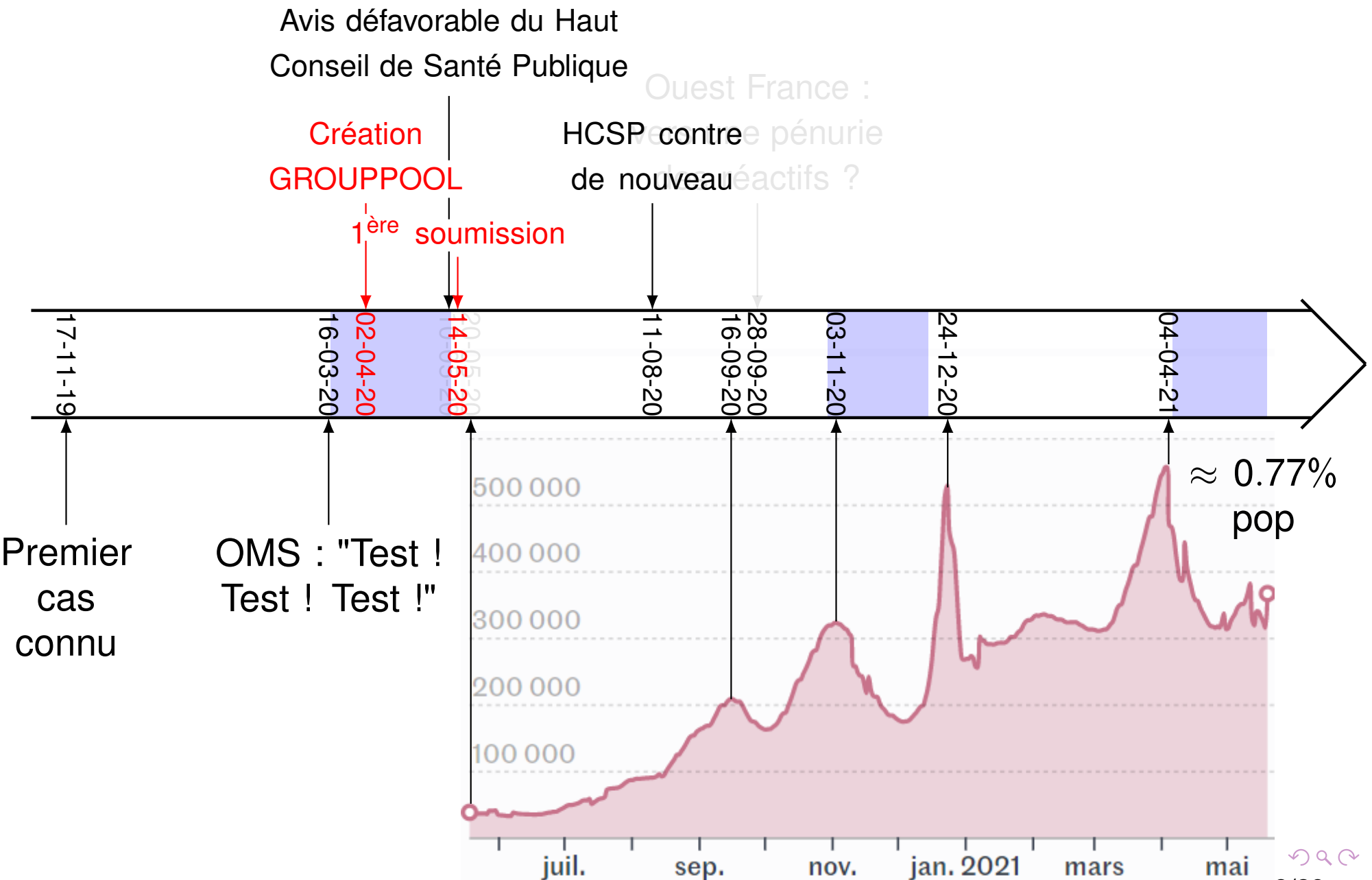
1. Utilisé par Dorfman [1943] pour détecter la syphilis chez les soldats américains : groupes de taille $n = 128$ puis deuxième test individuel de toutes les personnes appartenant à un groupe contaminé.
2. Utilisé en France pour détecter le VIH dans les produits du sang. Lié à l'affaire du sang contaminé ?
3. Envisagé pendant l'épidémie de SARS-COV-1 pour mettre en quarantaine régulièrement des groupes de patients suspects.
4. Création de Groupool le 2 avril 2020.

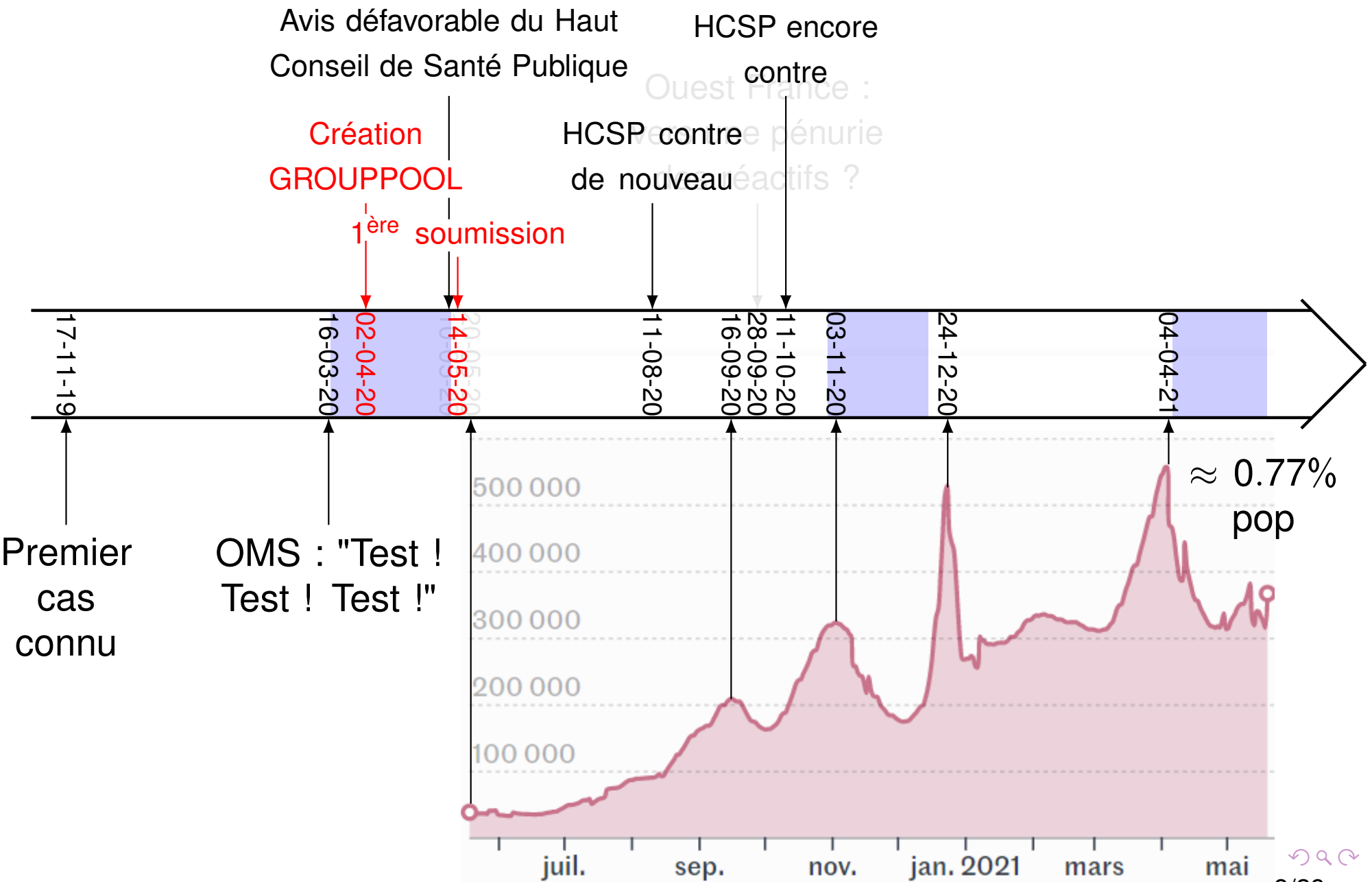

Ouest France :
vers une pénurie
des réactifs ?
17-11-19
16-03-20
20-05-20
16-09-20
28-09-20
03-11-20
24-12-20
04-04-21
Premier
cas
connu
OMS : "Test !
Test ! Test !"
≈ 0.77%
pop
500 000
400 000
300 000
200 000
100 000
juil.
sep.
nov.
jan. 2021
mars
mai

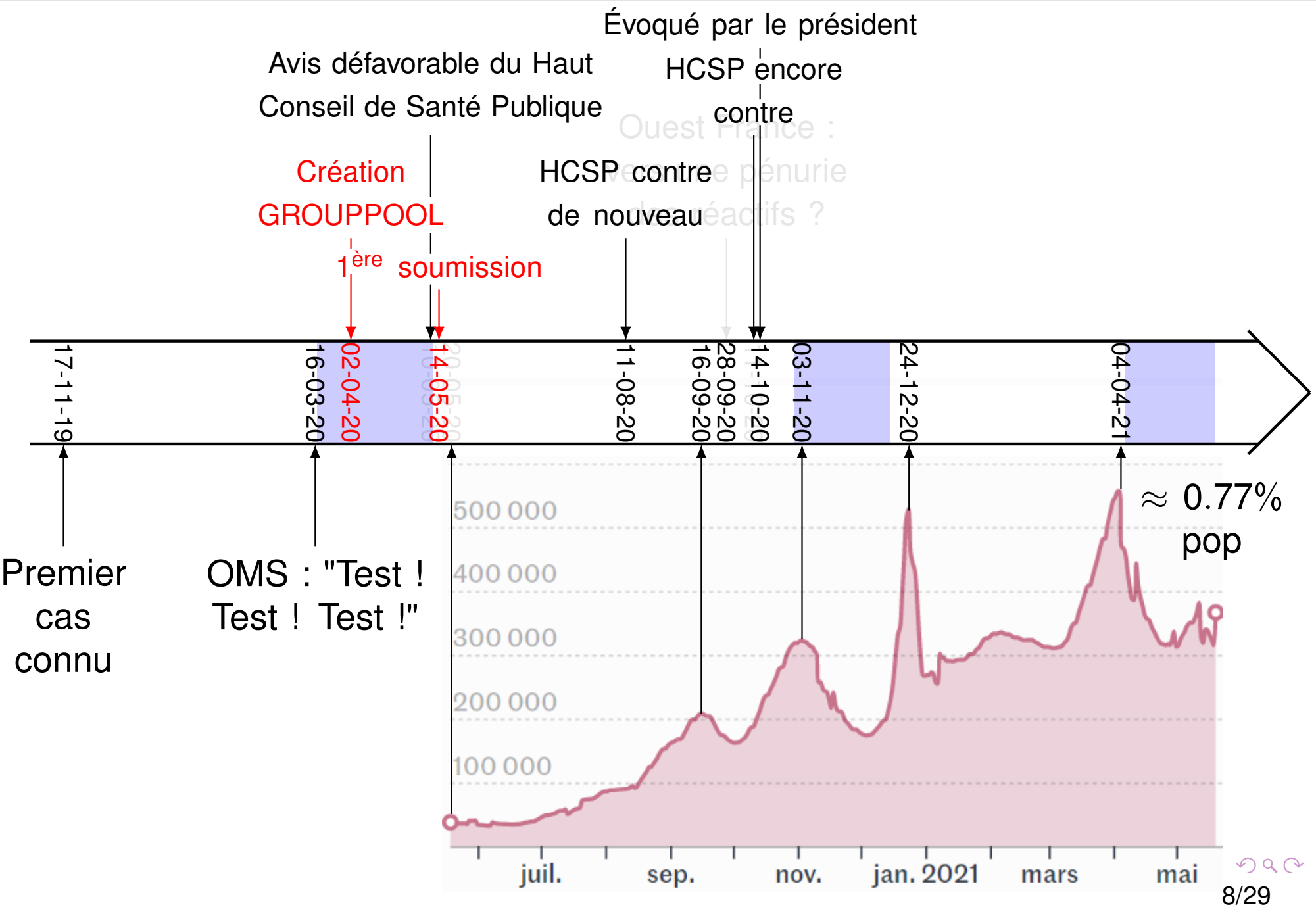

Création
GROUPPOOL
Ouest France :
vers une pénurie
des réactifs ?
17-11-19
16-03-20
02-04-20
20-05-20
16-09-20
28-09-20
03-11-20
24-12-20
04-04-21
Premier
cas
connu
OMS : "Test !
Test ! Test !"
≈ 0.77%
pop
500 000
400 000
300 000
200 000
100 000
juil.
sep.
nov.
jan. 2021
mars
mai

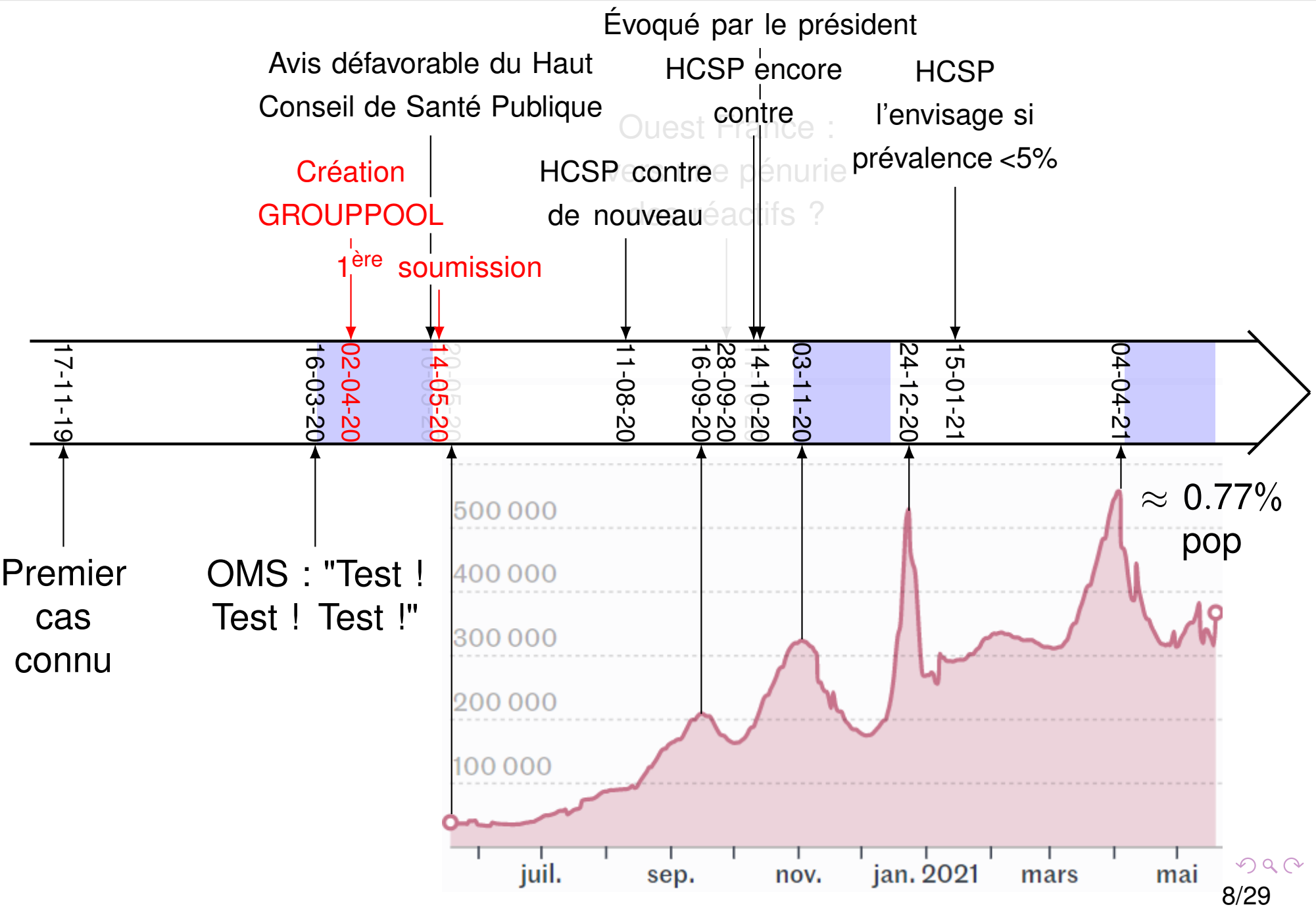

Avis défavorable du Haut Conseil de Santé Publique
Création GROUPPOOL
Ouest France : vers une pénurie des réactifs ?
17-11-19
16-03-20
02-04-20
20-05-20
10-05-20
28-09-20
16-09-20
03-11-20
24-12-20
04-04-21
Premier cas connu
OMS : "Test ! Test ! Test !"
≈ 0.77% pop
500 000
400 000
300 000
200 000
100 000
juil.
sep.
nov.
jan. 2021
mars
mai

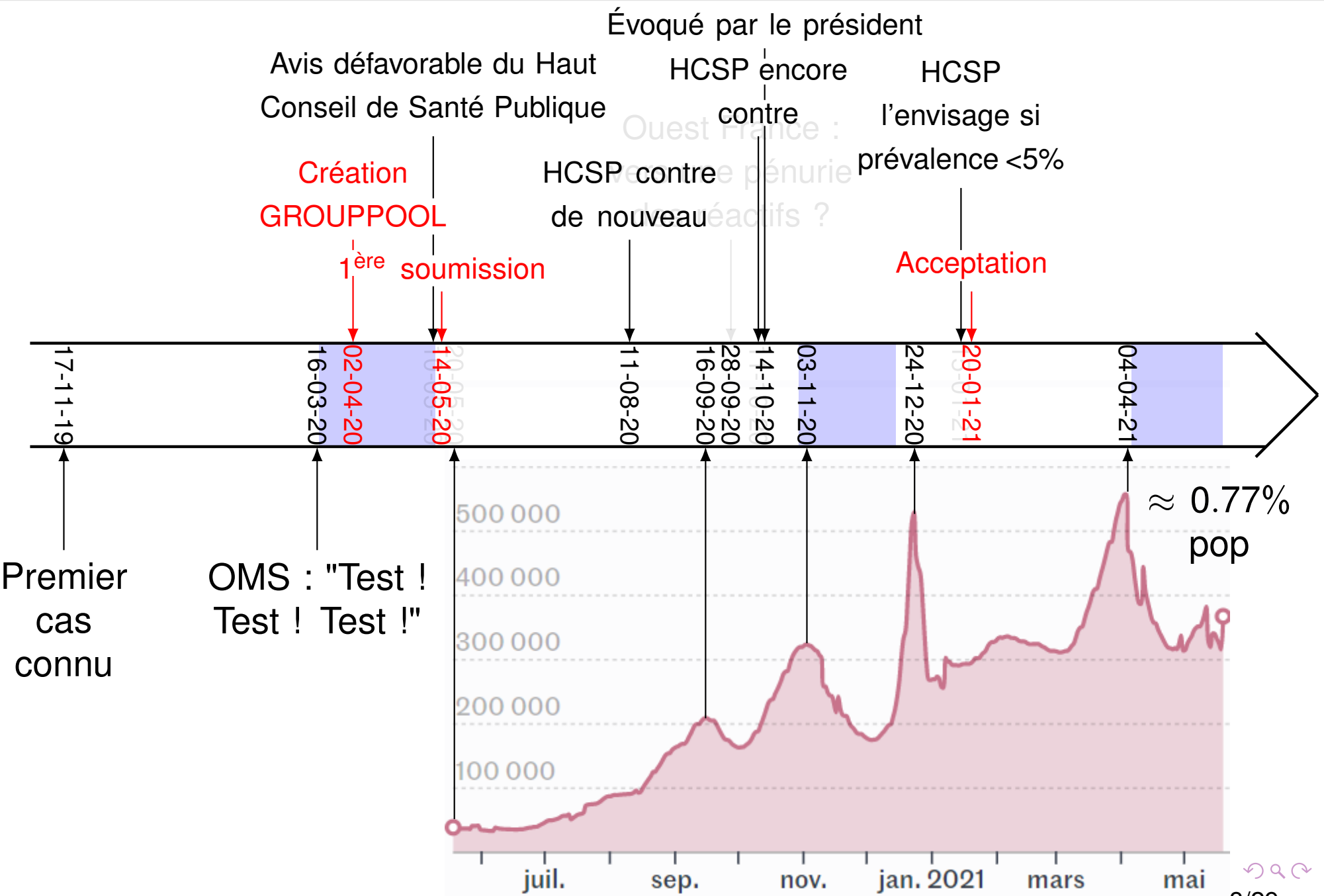

Avis défavorable du Haut Conseil de Santé Publique
Création GROUPPOOL
1ère soumission
Ouest France : vers une pénurie des réactifs ?
17-11-19
16-03-20
02-04-20
14-05-20
28-09-20
16-09-20
03-11-20
24-12-20
04-04-21
Premier cas connu
OMS : "Test ! Test ! Test !"
≈ 0.77% pop
500 000
400 000
300 000
200 000
100 000
juil.
sep.
nov.
jan. 2021
mars
mai


Avis défavorable du Haut Conseil de Santé Publique
Création GROUPPOOL
1ère soumission
HCSP contre de nouveau
17-11-19
16-03-20
02-04-20
14-05-20
11-08-20
16-09-20
28-09-20
03-11-20
24-12-20
04-04-21
Premier cas connu
OMS : "Test ! Test ! Test !"
≈ 0.77% pop
500 000
400 000
300 000
200 000
100 000
juil.
sep.
nov.
jan. 2021
mars
mai


Avis défavorable du Haut Conseil de Santé Publique
HCSP encore contre
Création GROUPPOOL
1ère soumission
HCSP contre de nouveau
17-11-19
16-03-20
02-04-20
14-05-20
11-08-20
16-09-20
28-09-20
11-10-20
03-11-20
24-12-20
04-04-21
Premier cas connu
OMS : "Test ! Test ! Test !"
≈ 0.77% pop
500 000
400 000
300 000
200 000
100 000
juil.
sep.
nov.
jan. 2021
mars
mai


Évoqué par le président
Avis défavorable du Haut Conseil de Santé Publique
HCSP encore contre
Création GROUPPOOL
HCSP contre de nouveau
1ère soumission
17-11-19
16-03-20
02-04-20
14-05-20
11-08-20
16-09-20
28-09-20
14-10-20
03-11-20
24-12-20
04-04-21
Premier cas connu
OMS : "Test ! Test ! Test !"
≈ 0.77% pop
500 000
400 000
300 000
200 000
100 000
juil.
sep.
nov.
jan. 2021
mars
mai


Évoqué par le président
Avis défavorable du Haut Conseil de Santé Publique
HCSP encore contre
HCSP l'envisage si prévalence <5%
Création GROUPPOOL
1ère soumission
HCSP contre de nouveau
17-11-19
16-03-20
02-04-20
14-05-20
11-08-20
16-09-20
28-09-20
14-10-20
03-11-20
24-12-20
15-01-21
04-04-21
Premier cas connu
OMS : "Test ! Test ! Test !"
≈ 0.77% pop
500 000
400 000
300 000
200 000
100 000
juil.
sep.
nov.
jan. 2021
mars
mai


Évoqué par le président
Avis défavorable du Haut Conseil de Santé Publique
HCSP encore contre
HCSP l'envisage si prévalence <5%
Création GROUPPOOL
HCSP contre de nouveau
1ère soumission
Acceptation
17-11-19
16-03-20
02-04-20
14-05-20
11-08-20
16-09-20
28-09-20
14-10-20
03-11-20
24-12-20
20-01-21
04-04-21
Premier cas connu
OMS : "Test ! Test ! Test !"
≈ 0.77% pop
500 000
400 000
300 000
200 000
100 000
juil.
sep.
nov.
jan. 2021
mars
mai

Le pooling dans le monde pour le SARS-COV-2

02-20 Aux États-Unis, dans la baie de San Francisco, utilisation pour estimer la prévalence.

Le pooling dans le monde pour le SARS-COV-2

02-20 Aux États-Unis, dans la baie de San Francisco, utilisation pour estimer la prévalence.

05-03-20 L'Allemagne fait des tests groupés dans des structures hospitalières et des maisons de retraites à des fins de préventions épidémiques (comme au Portugal et à Singapour).

Le pooling dans le monde pour le SARS-COV-2

02-20 Aux États-Unis, dans la baie de San Francisco, utilisation pour estimer la prévalence.

05-03-20 L'Allemagne fait des tests groupés dans des structures hospitalières et des maisons de retraites à des fins de préventions épidémiques (comme au Portugal et à Singapour).

01-06-20 En Inde, le pooling est recommandé comme surveillance pour les zones avec un taux de positivité inférieur à 5%.

Le pooling dans le monde pour le SARS-COV-2

02-20 Aux États-Unis, dans la baie de San Francisco, utilisation pour estimer la prévalence.

05-03-20 L'Allemagne fait des tests groupés dans des structures hospitalières et des maisons de retraites à des fins de préventions épidémiques (comme au Portugal et à Singapour).

01-06-20 En Inde, le pooling est recommandé comme surveillance pour les zones avec un taux de positivité inférieur à 5%.

23-06-20 En Israël, Ben-Ami et al. [2020] montrent que le pooling accroît les capacités de test en maintenant un haut degrés de sensitivité.

Le pooling dans le monde pour le SARS-COV-2

02-20 Aux États-Unis, dans la baie de San Francisco, utilisation pour estimer la prévalence.

05-03-20 L'Allemagne fait des tests groupés dans des structures hospitalières et des maisons de retraites à des fins de préventions épidémiques (comme au Portugal et à Singapour).

01-06-20 En Inde, le pooling est recommandé comme surveillance pour les zones avec un taux de positivité inférieur à 5%.

23-06-20 En Israël, Ben-Ami et al. [2020] montrent que le pooling accroît les capacités de test en maintenant un haut degrés de sensitivité.

07-20 Le Rwanda utilise le pooling pour diminuer les coûts notamment (voir Mutesa et al. [2020]).

Le pooling dans le monde pour le SARS-COV-2

02-20 Aux États-Unis, dans la baie de San Francisco, utilisation pour estimer la prévalence.

05-03-20 L'Allemagne fait des tests groupés dans des structures hospitalières et des maisons de retraites à des fins de préventions épidémiques (comme au Portugal et à Singapour).

01-06-20 En Inde, le pooling est recommandé comme surveillance pour les zones avec un taux de positivité inférieur à 5%.

23-06-20 En Israël, Ben-Ami et al. [2020] montrent que le pooling accroît les capacités de test en maintenant un haut degrés de sensitivité.

07-20 Le Rwanda utilise le pooling pour diminuer les coûts notamment (voir Mutesa et al. [2020]).

15-10-20 En Israël, Yelin et al. [2020] montrent l'efficacité du pooling.

Le pooling dans le monde pour le SARS-COV-2

02-20 Aux États-Unis, dans la baie de San Francisco, utilisation pour estimer la prévalence.

05-03-20 L'Allemagne fait des tests groupés dans des structures hospitalières et des maisons de retraites à des fins de préventions épidémiques (comme au Portugal et à Singapour).

01-06-20 En Inde, le pooling est recommandé comme surveillance pour les zones avec un taux de positivité inférieur à 5%.

23-06-20 En Israël, Ben-Ami et al. [2020] montrent que le pooling accroît les capacités de test en maintenant un haut degrés de sensitivité.

07-20 Le Rwanda utilise le pooling pour diminuer les coûts notamment (voir Mutesa et al. [2020]).

15-10-20 En Israël, Yelin et al. [2020] montrent l'efficacité du pooling.

- *Chai Bio* vend *OPEN QPCR* qui permet de faire du pooling salivaire et d'avoir une réponse en deux heures.

Plan

Revenons à la réalité

Un test parfait n'existe pas. Quelles conséquences sur le pooling ?

Revenons à la réalité

Un test parfait n'existe pas. Quelles conséquences sur le pooling ?

- Un faux positif implique qu'on teste à nouveau tout un groupe pour rien.

Revenons à la réalité

Un test parfait n'existe pas. Quelles conséquences sur le pooling ?

- Un faux positif implique qu'on teste à nouveau tout un groupe pour rien.
- Un faux négatif implique qu'on rate des contaminés.

Revenons à la réalité

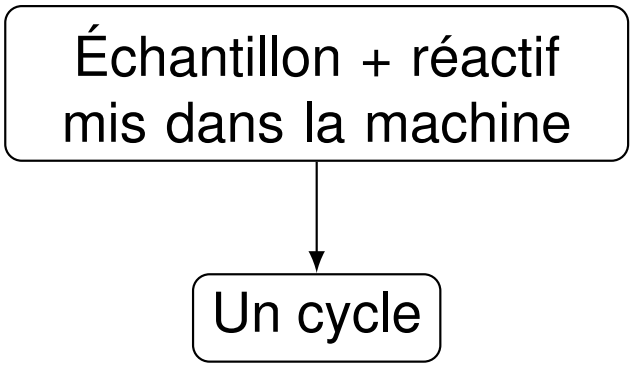
Un test parfait n'existe pas. Quelles conséquences sur le pooling ?

- Un faux positif implique qu'on teste à nouveau tout un groupe pour rien.
- Un faux négatif implique qu'on rate des contaminés.
  $\hookrightarrow$ dilution $\Rightarrow$ faux négatif ?

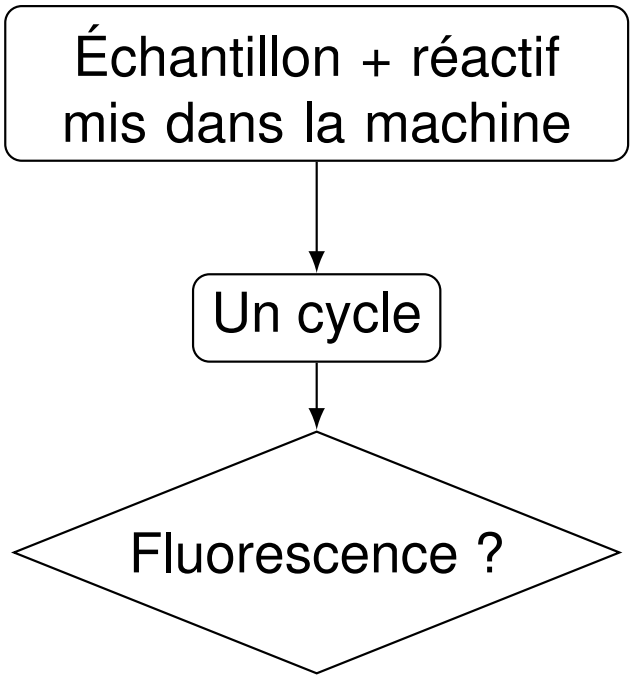
Test PCR

Échantillon + réactif
mis dans la machine

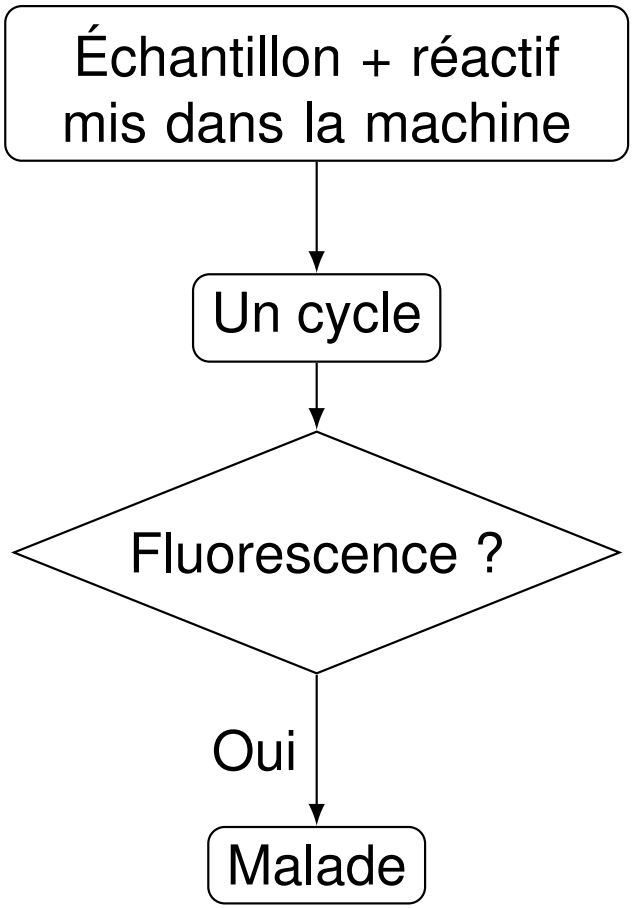
Test PCR



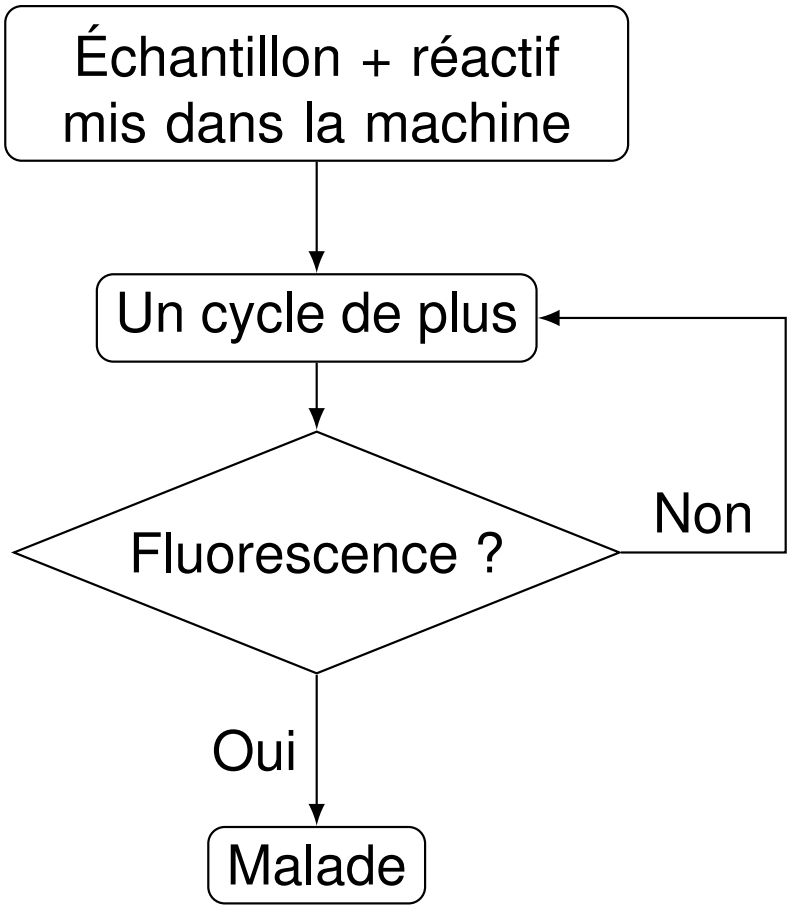
Test PCR


Échantillon + réactif
mis dans la machine
Un cycle
Fluorescence ?

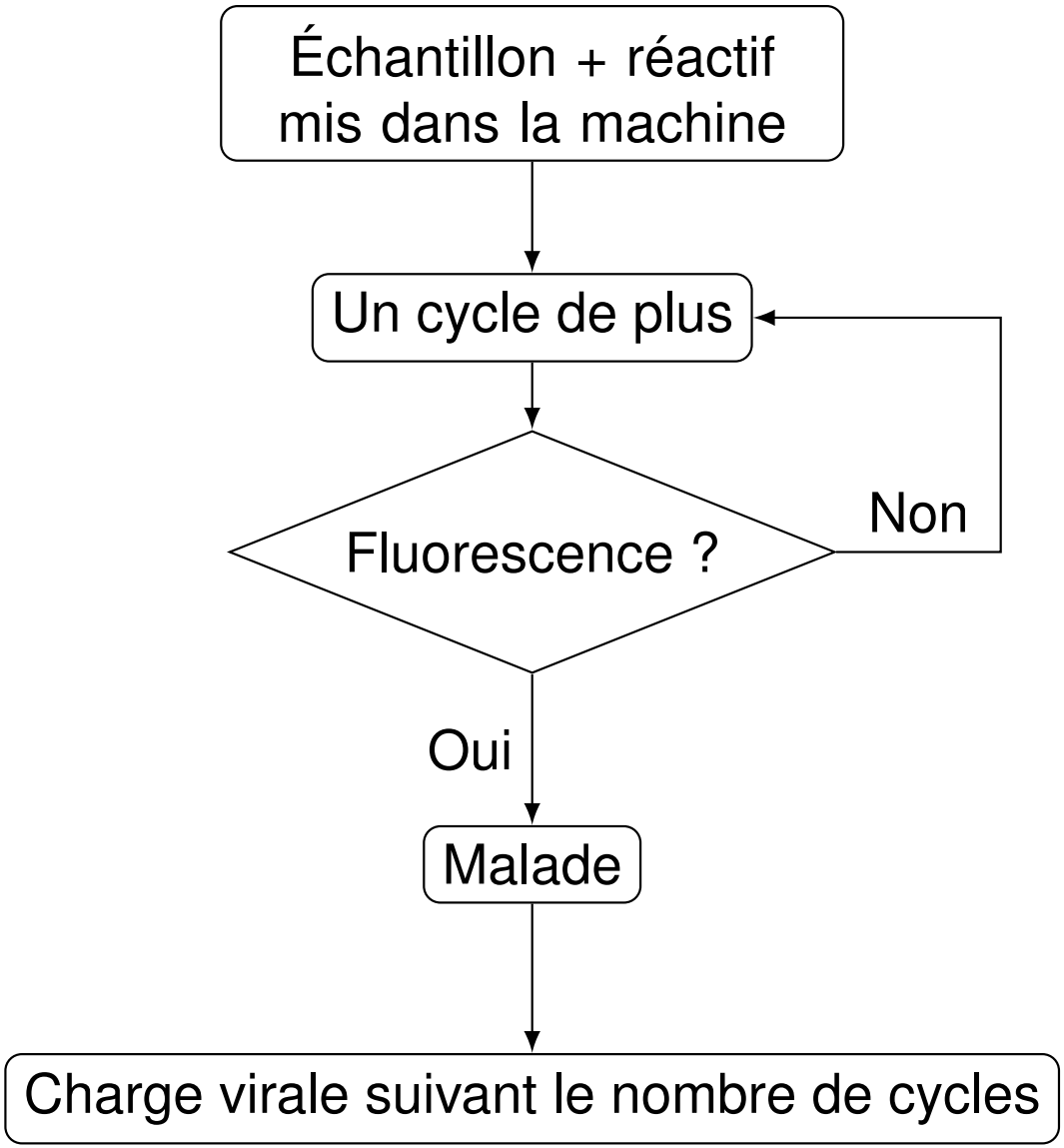
Test PCR


Échantillon + réactif
mis dans la machine
Un cycle
Fluorescence ?
Oui
Malade

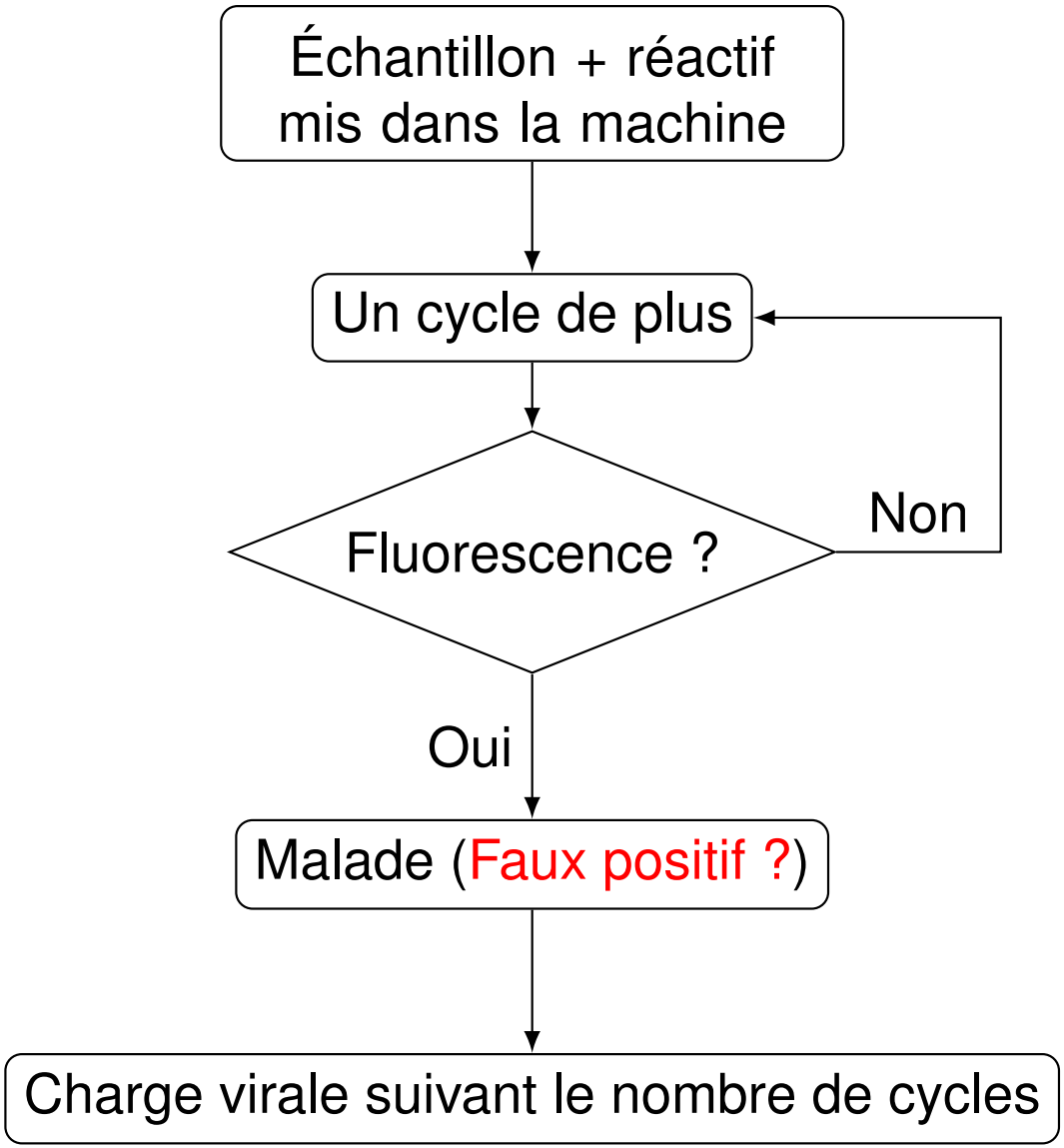
Test PCR


Échantillon + réactif
mis dans la machine
Un cycle de plus
Fluorescence ?
Non
Oui
Malade

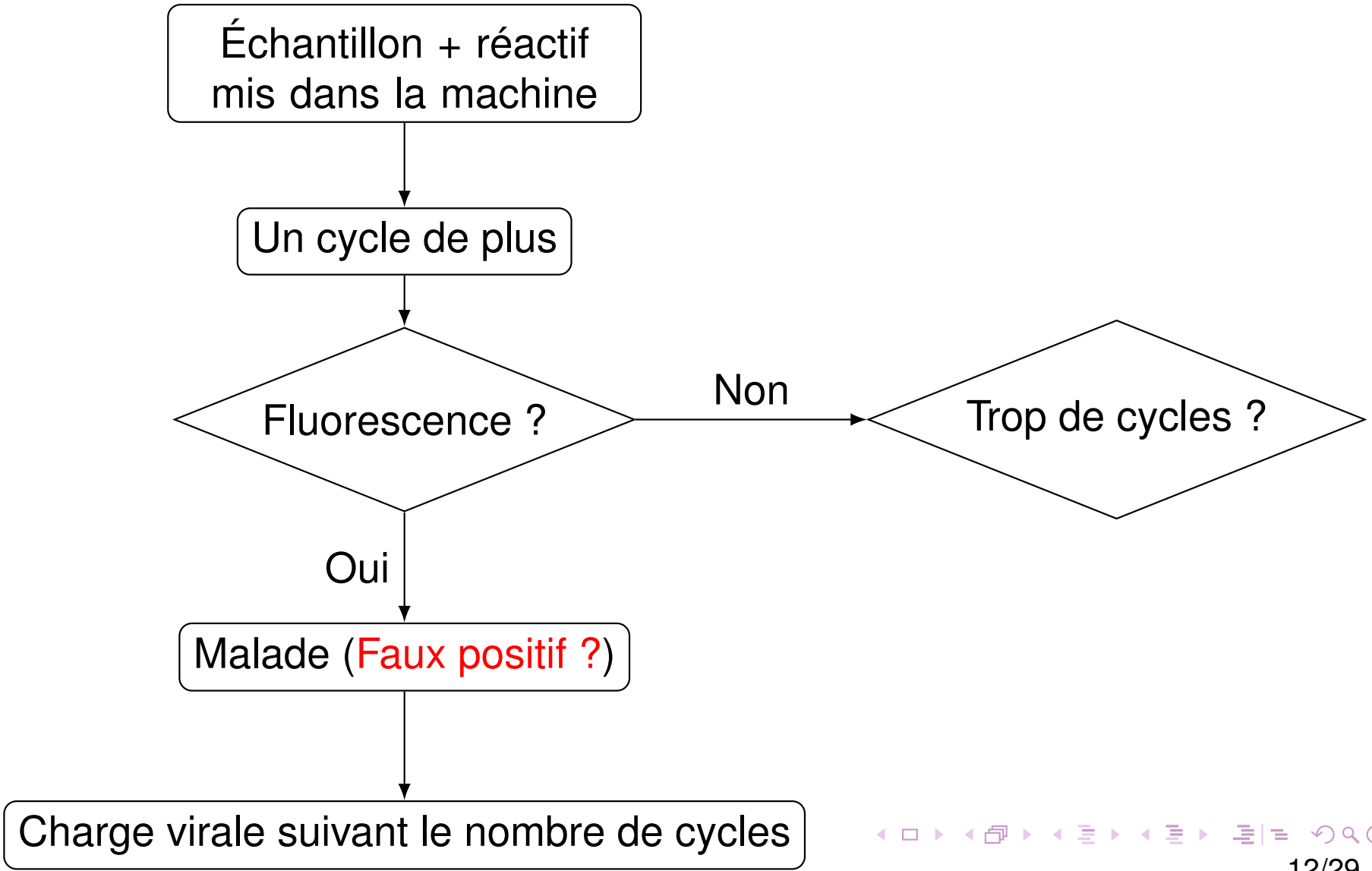
Test PCR


Échantillon + réactif
mis dans la machine
Un cycle de plus
Fluorescence ?
Non
Oui
Malade
Charge virale suivant le nombre de cycles

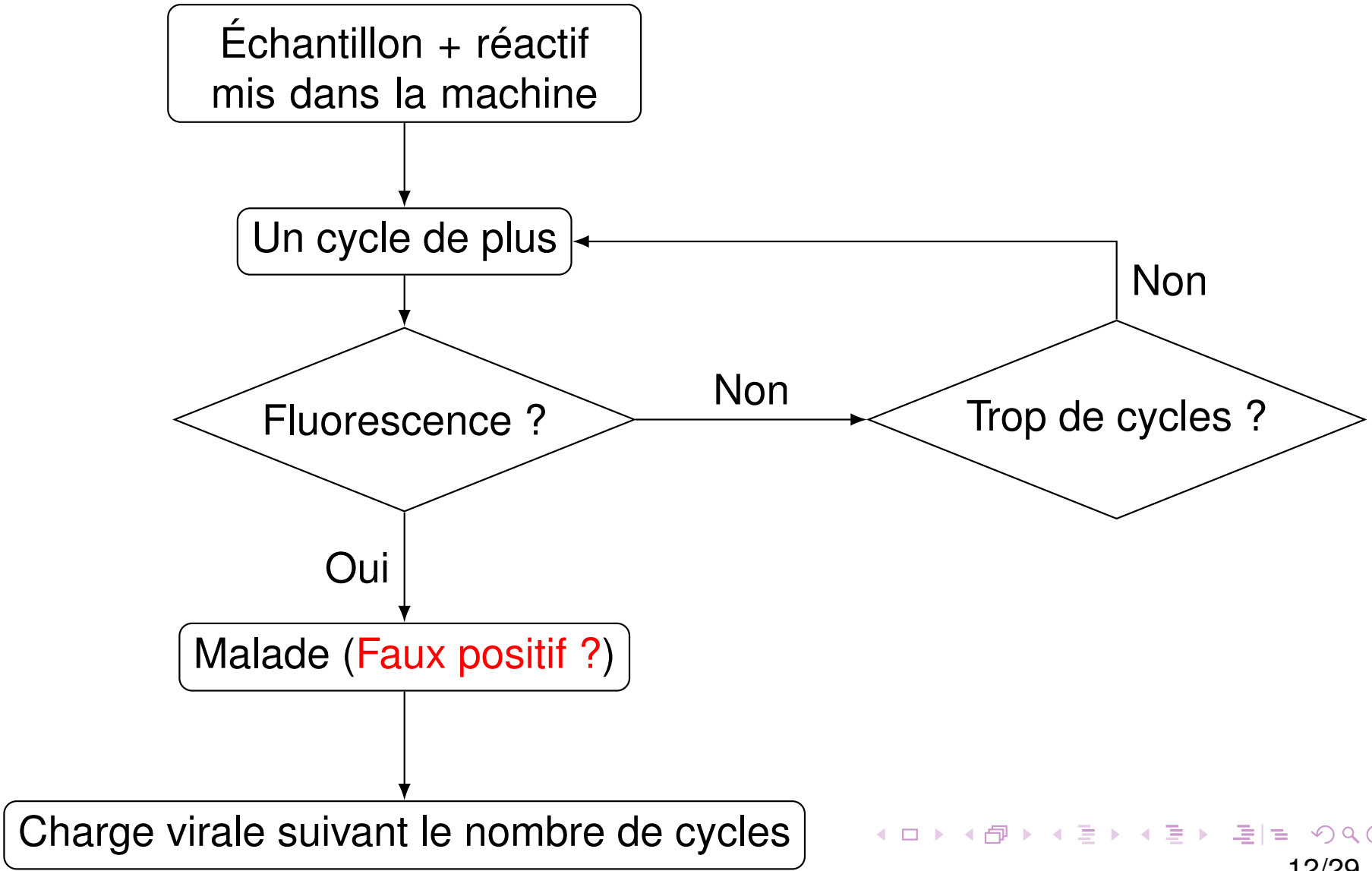
Test PCR


Échantillon + réactif
mis dans la machine
Un cycle de plus
Fluorescence ?
Non
Oui
Malade (Faux positif ?)
Charge virale suivant le nombre de cycles

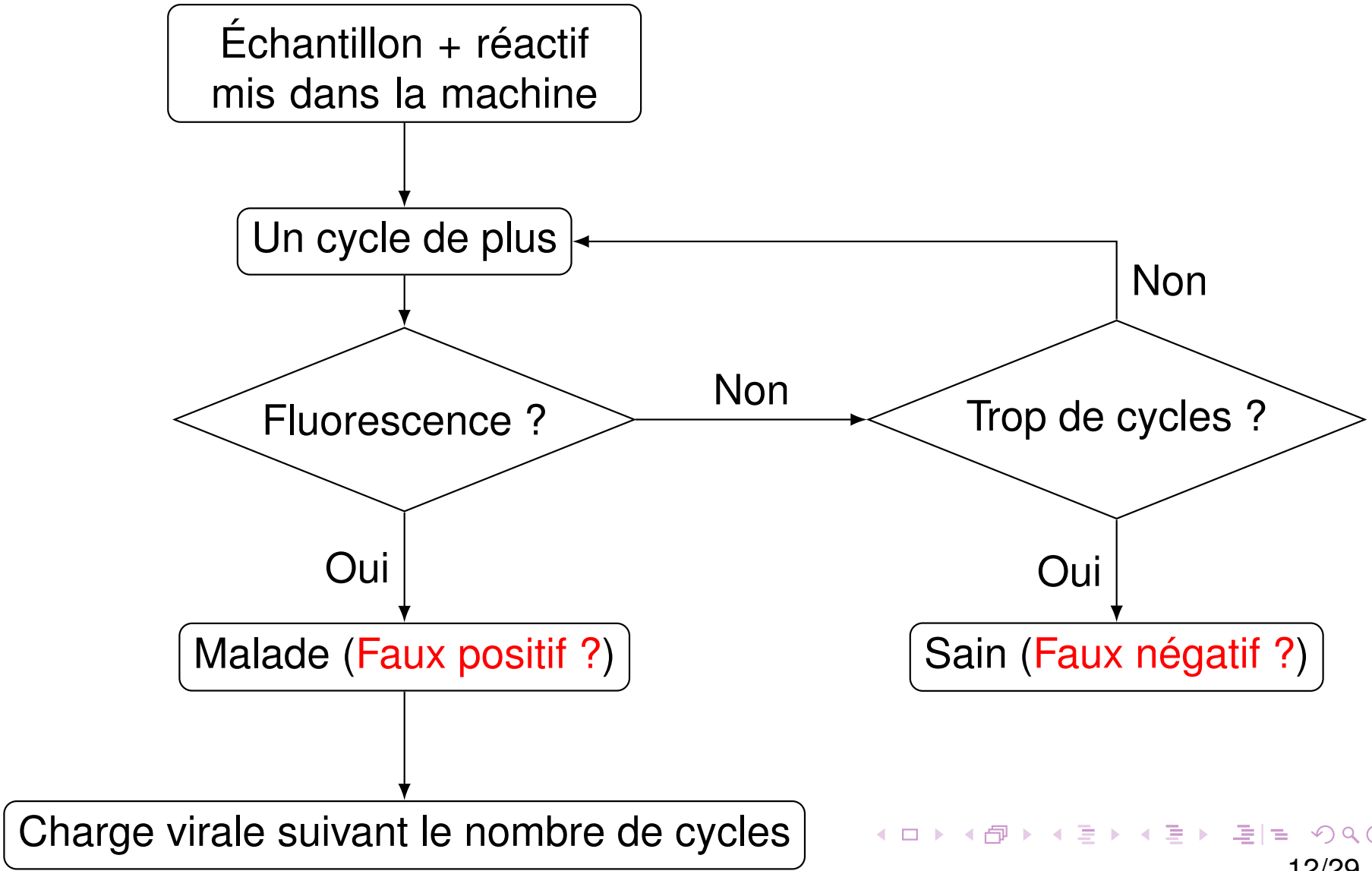
Test PCR


Échantillon + réactif
mis dans la machine
Un cycle de plus
Fluorescence ?
Non
Trop de cycles ?
Oui
Malade (Faux positif ?)
Charge virale suivant le nombre de cycles

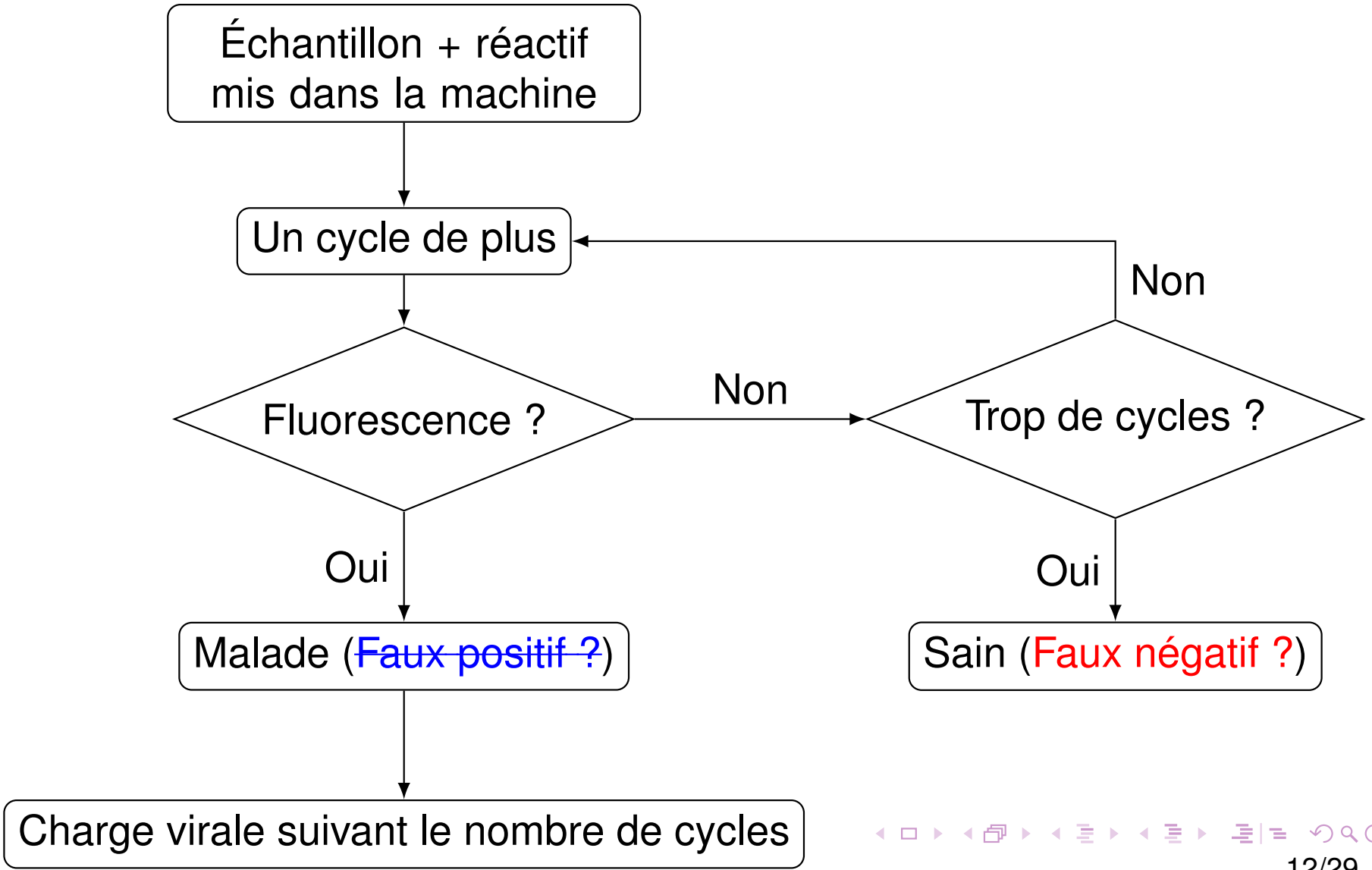
Test PCR


Échantillon + réactif
mis dans la machine
Un cycle de plus
Fluorescence ?
Non
Trop de cycles ?
Non
Oui
Malade (Faux positif ?)
Charge virale suivant le nombre de cycles

Test PCR


Échantillon + réactif
mis dans la machine
Un cycle de plus
Fluorescence ?
Non
Trop de cycles ?
Non
Oui
Oui
Malade (Faux positif ?)
Sain (Faux négatif ?)
Charge virale suivant le nombre de cycles

Test PCR


Échantillon + réactif
mis dans la machine
Un cycle de plus
Fluorescence ?
Non
Trop de cycles ?
Non
Oui
Oui
Malade (Faux positif ?)
Sain (Faux négatif ?)
Charge virale suivant le nombre de cycles

Modélisation : un individu

Pour un patient i, notons Y_i le nombre de cycles (sans censure) alors :

$$Y_i = \quad - \log_2 \quad C_i$$

où C_i est la charge virale du patient,

Modélisation : un individu

Pour un patient i, notons Y_i le nombre de cycles (sans censure) alors :

$$Y_i = \qquad -\log_2\left(C_i + \varepsilon_i^{(1)}\right)$$

où C_i est la charge virale du patient, $\varepsilon_i^{(1)}$ suit une loi log-normale de paramètres (ν, τ^2) et représente le bruit résiduel impliquant un potentiel faux positif,

Modélisation : un individu

Pour un patient i, notons Y_i le nombre de cycles (sans censure) alors :

$$Y_i = \quad - \log_2 \left(C_i + \varepsilon_i^{(1)} \right) \quad + \varepsilon_i^{(2)}$$

où C_i est la charge virale du patient, $\varepsilon_i^{(1)}$ suit une loi log-normale de paramètres (ν, τ^2) et représente le bruit résiduel impliquant un potentiel faux positif, $\varepsilon_i^{(2)}$ suit une loi gaussienne centrée de variance ρ^2 et est le bruit de mesure intrinsèque à la machine

Modélisation : un individu

Pour un patient i, notons Y_i le nombre de cycles (avec censure) alors :

$$Y_i = \min \left[-\log_2 \left(C_i + \varepsilon_i^{(1)} \right), d_{cens} \right] + \varepsilon_i^{(2)}$$

où C_i est la charge virale du patient, $\varepsilon_i^{(1)}$ suit une loi log-normale de paramètres (ν, τ^2) et représente le bruit résiduel impliquant un potentiel faux positif, $\varepsilon_i^{(2)}$ suit une loi gaussienne centrée de variance ρ^2 et est le bruit de mesure intrinsèque à la machine et d_{cens} le seuil limite des cycles.

Modélisation : un individu

Pour un patient i, notons Y_i le nombre de cycles (avec censure) alors :

$$Y_i = \min \left[-\log_2 \left(C_i + \varepsilon_i^{(1)} \right), d_{cens} \right] + \varepsilon_i^{(2)}$$

où C_i est la charge virale du patient, $\varepsilon_i^{(1)}$ suit une loi log-normale de paramètres (ν, τ^2) et représente le bruit résiduel impliquant un potentiel faux positif, $\varepsilon_i^{(2)}$ suit une loi gaussienne centrée de variance ρ^2 et est le bruit de mesure intrinsèque à la machine et d_{cens} le seuil limite des cycles.

Nous supposons que $\mathbb{P}\left(\varepsilon_i^{(1)} > 2^{-d_{cens}}\right) << 1$

Modélisation : un individu

Pour un patient i, notons Y_i le nombre de cycles (avec censure) alors :

$$Y_i = \min\left[-\log_2\left(C_i + \varepsilon_i^{(1)}\right), d_{cens}\right] + \varepsilon_i^{(2)}$$

où C_i est la charge virale du patient, $\varepsilon_i^{(1)}$ suit une loi log-normale de paramètres (ν, τ^2) et représente le bruit résiduel impliquant un potentiel faux positif, $\varepsilon_i^{(2)}$ suit une loi gaussienne centrée de variance ρ^2 et est le bruit de mesure intrinsèque à la machine et d_{cens} le seuil limite des cycles.

Nous supposons que $\mathbb{P}\left(\varepsilon_i^{(1)} > 2^{-d_{cens}}\right) << 1$ donc, comme $\log(a+b) \approx \log[\max(a,b)]$, alors :

$$Y_i \approx \min\left[-\log_2\left(C_i\right), d_{cens}\right] + \varepsilon_i^{(2)}.$$

Modélisation : pooling

Pour la modélisation du pooling, nous prenons N individus avec des charges virales $C_1, \ldots, C_N$ positives telles que $\mathbb{P}(C_i > 0) = p$ où p est la prévalence de la population.

Modélisation : pooling

Pour la modélisation du pooling, nous prenons N individus avec des charges virales $C_1, \ldots, C_N$ positives telles que $\mathbb{P}(C_i > 0) = p$ où p est la prévalence de la population. Après mélange des échantillons, nous supposons que la charge virale est donc :

$$C^{(N)} \approx \frac{1}{N} \sum_{j=1}^{N} C_j$$

Modélisation : pooling

Pour la modélisation du pooling, nous prenons N individus avec des charges virales $C_1, \ldots, C_N$ positives telles que $\mathbb{P}(C_i > 0) = p$ où p est la prévalence de la population. Après mélange des échantillons, nous supposons que la charge virale est donc :

$$C^{(N)} \approx \frac{1}{N} \sum_{j=1}^{N} C_j$$

et le nombre de cycles de l'échantillon du groupe est donc :

$$Y_j^{(N)} \approx \min \left[-\log_2 \left(\frac{1}{N} \sum_{j=1}^{N} C_j \right), d_{cens} \right] + \varepsilon_i^{(2)}$$

Modélisation : pooling

Pour la modélisation du pooling, nous prenons N individus avec des charges virales $C_1, \ldots, C_N$ positives telles que $\mathbb{P}(C_i > 0) = p$ où p est la prévalence de la population. Après mélange des échantillons, nous supposons que la charge virale est donc :

$$C^{(N)} \approx \frac{1}{N} \sum_{j=1}^{N} C_j$$

et le nombre de cycles de l'échantillon du groupe est donc :

$$\begin{aligned} Y_j^{(N)} &\approx \min \left[-\log_2 \left(\frac{1}{N} \sum_{j=1}^{N} C_j \right), d_{cens} \right] + \varepsilon_i^{(2)} \\ &\approx \min \left[\log_2(N) - \log_2 \left(\sum_{j=1}^{N} C_j \right), d_{cens} \right] + \varepsilon_i^{(2)} \end{aligned}$$

Modélisation : pooling

Pour la modélisation du pooling, nous prenons N individus avec des charges virales $C_1, \ldots, C_N$ positives telles que $\mathbb{P}(C_i > 0) = p$ où p est la prévalence de la population. Après mélange des échantillons, nous supposons que la charge virale est donc :

$$C^{(N)} \approx \frac{1}{N} \sum_{j=1}^{N} C_j$$

et le nombre de cycles de l'échantillon du groupe est donc :

$$\begin{aligned} Y_j^{(N)} &\approx \min \left[-\log_2 \left(\frac{1}{N} \sum_{j=1}^{N} C_j \right), d_{cens} \right] + \varepsilon_i^{(2)} \\ &\approx \min \left[\log_2(N) - \log_2 \left(\sum_{j=1}^{N} C_j \right), d_{cens} \right] + \varepsilon_i^{(2)} \\ &\approx \min \left[\log_2(N) - \log_2 \left(\max_{j=1,\ldots,N} C_j \right), d_{cens} \right] + \varepsilon_i^{(2)} \end{aligned}$$

Modélisation : pooling

Pour la modélisation du pooling, nous prenons N individus avec des charges virales $C_1, \ldots, C_N$ positives telles que $\mathbb{P}(C_i > 0) = p$ où p est la prévalence de la population. Après mélange des échantillons, nous supposons que la charge virale est donc :

$$C^{(N)} \approx \frac{1}{N} \sum_{j=1}^{N} C_j$$

et le nombre de cycles de l'échantillon du groupe est donc :

$$\begin{aligned} Y_j^{(N)} &\approx \min \left[-\log_2 \left(\frac{1}{N} \sum_{j=1}^{N} C_j \right), d_{cens} \right] + \varepsilon_i^{(2)} \\ &\approx \min \left[\log_2(N) - \log_2 \left(\sum_{j=1}^{N} C_j \right), d_{cens} \right] + \varepsilon_i^{(2)} \\ &\approx \min \left[\log_2(N) - \log_2 \left(\max_{j=1,\ldots,N} C_j \right), d_{cens} \right] + \varepsilon_i^{(2)} \end{aligned}$$

Plan

Quelle est la distribution du nombre de cycles ?

$\hookrightarrow$ pas du tout d'open data...

Quelle est la distribution du nombre de cycles ?

↪ pas du tout d'open data...

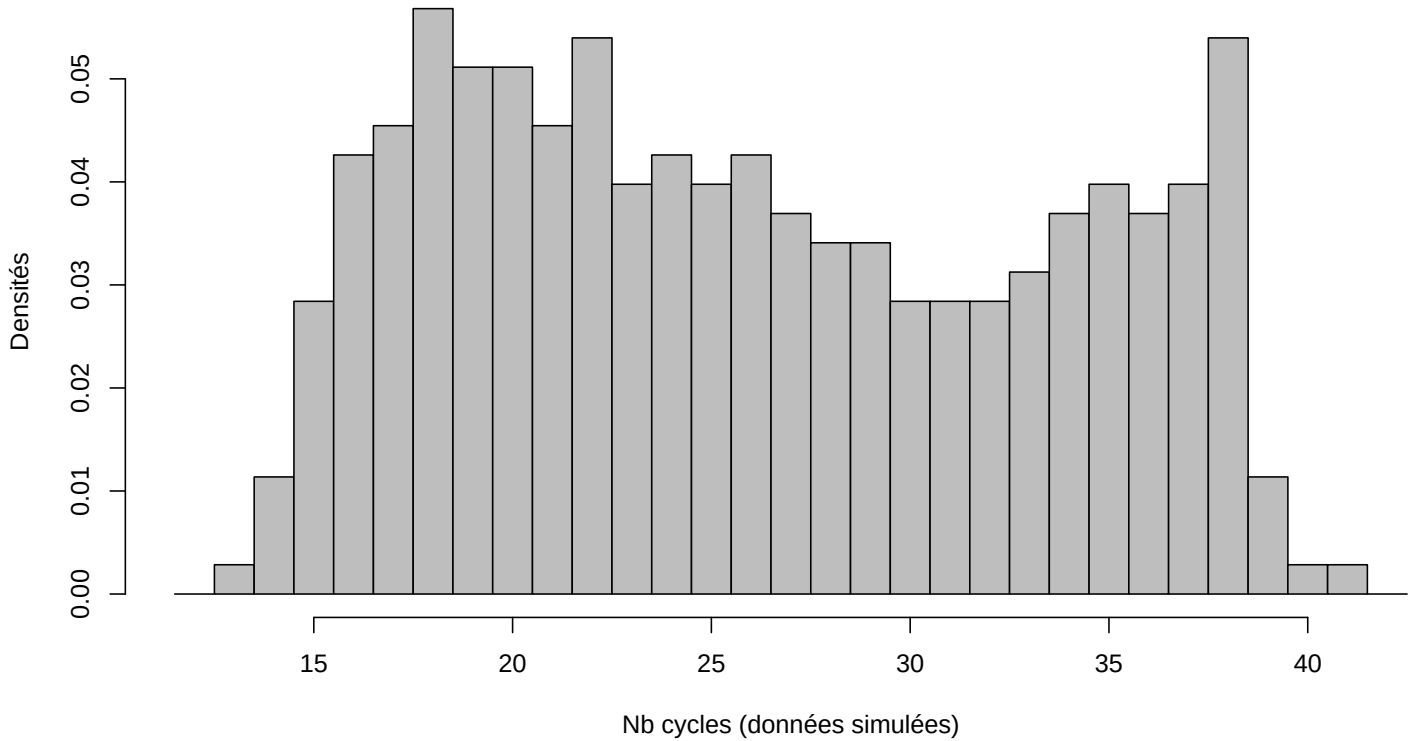


Figure: Reconstitution de l'histogramme proposé par Jones et al. [2020] avec 3 712 individus.

Quelle est la distribution du nombre de cycles ?

↪ pas du tout d'open data...

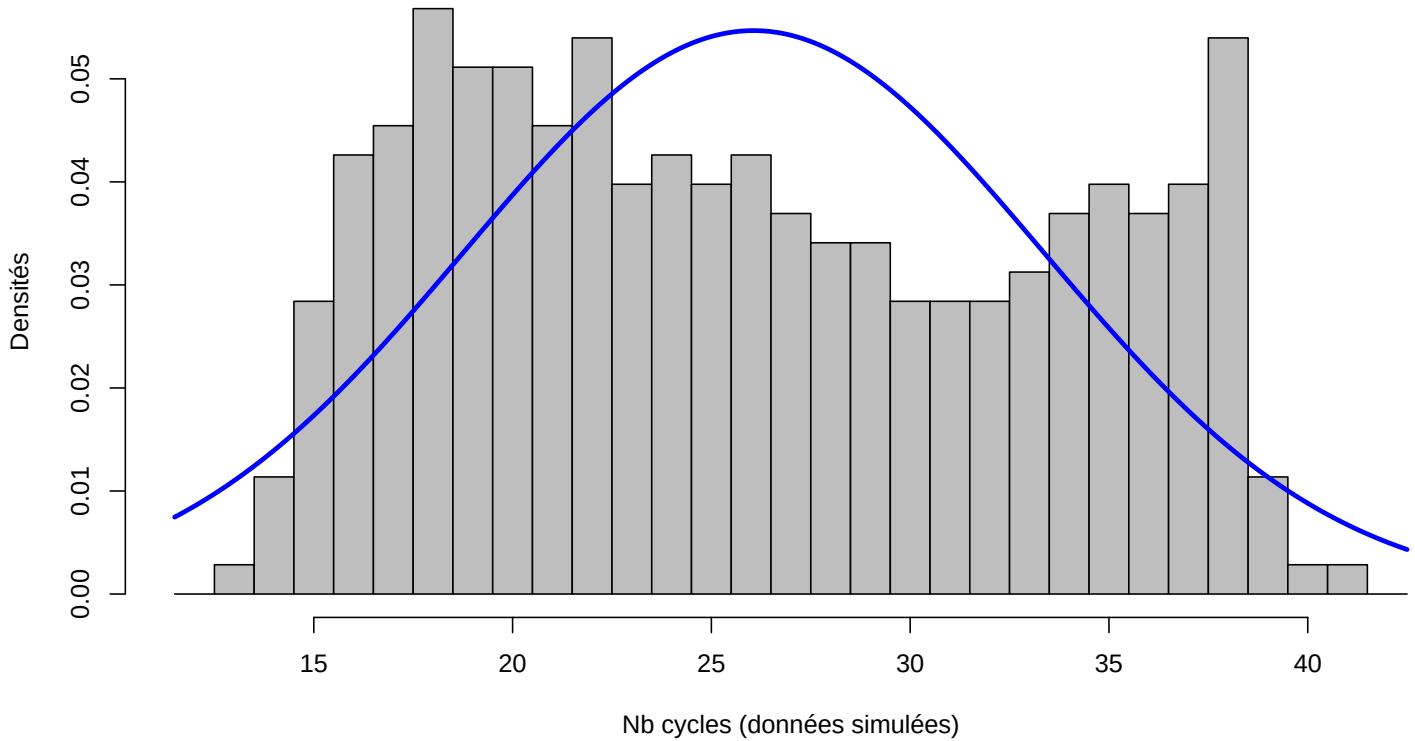


Figure: Reconstitution de l'histogramme proposé par Jones et al. [2020] avec 3 712 individus. Estimation avec un mélange d'une loi.

Quelle est la distribution du nombre de cycles ?

↪ pas du tout d'open data...

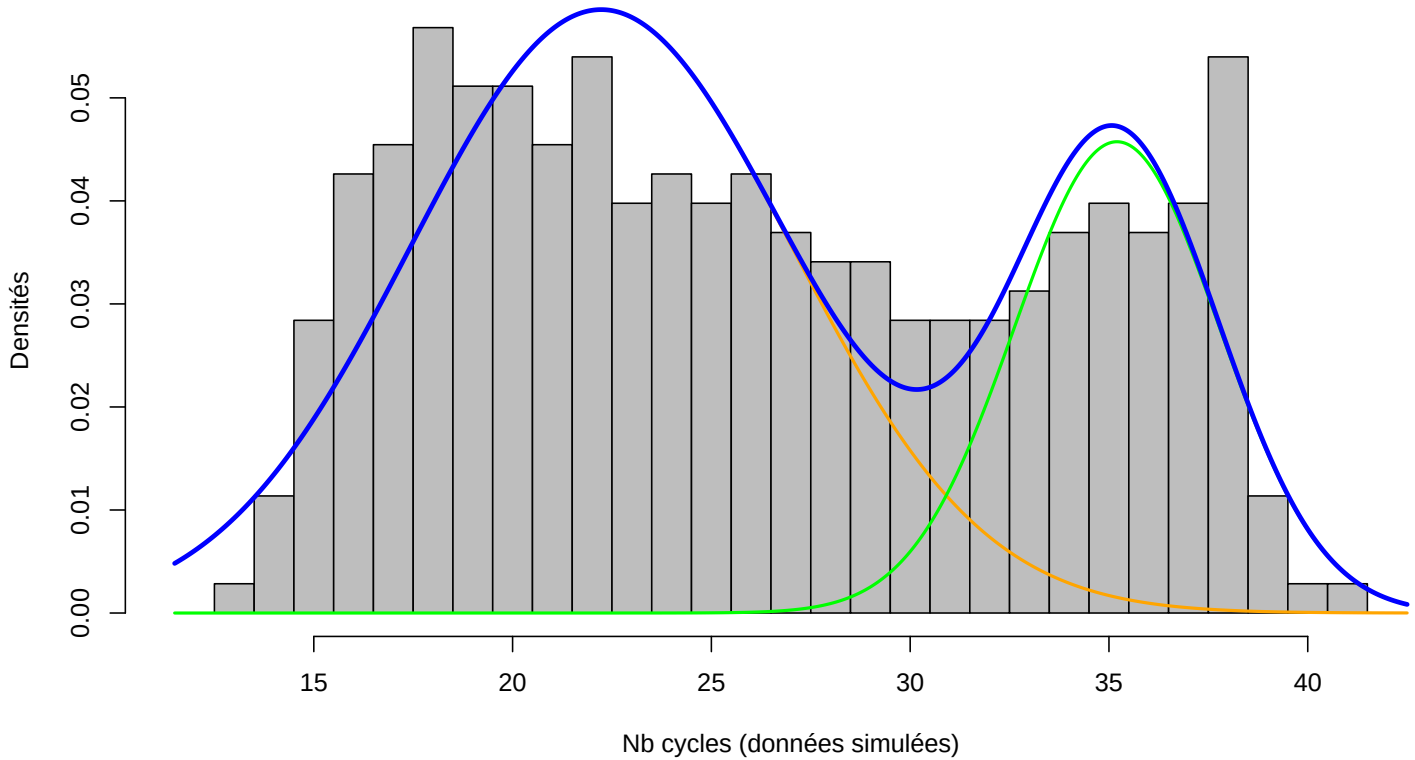


Figure: Reconstitution de l'histogramme proposé par Jones et al. [2020] avec 3 712 individus. Estimation avec un mélange de deux lois.

Quelle est la distribution du nombre de cycles ?

↪ pas du tout d'open data...

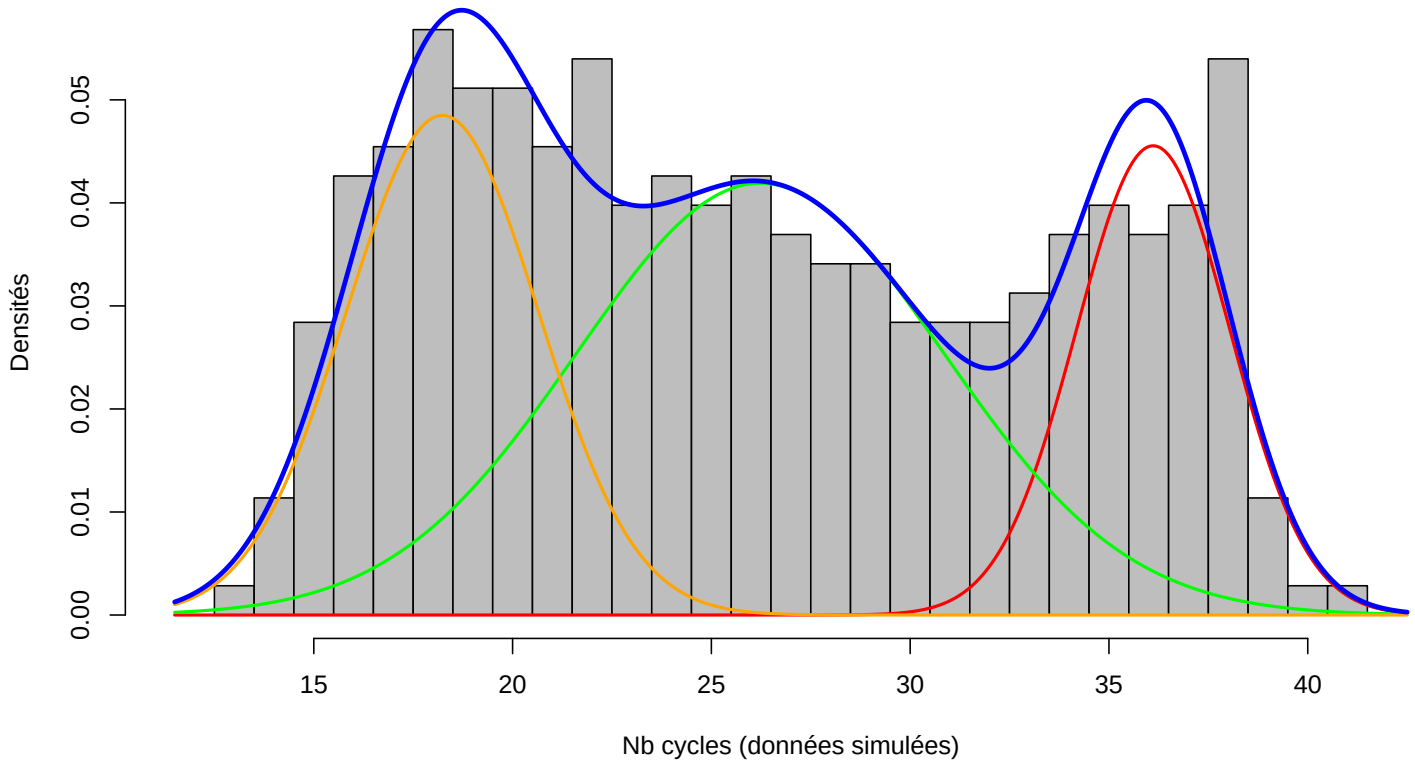


Figure: Reconstitution de l'histogramme proposé par Jones et al. [2020] avec 3 712 individus. Estimation avec un mélange de trois lois.

Quelle est la distribution du nombre de cycles ?

↪ pas du tout d'open data...

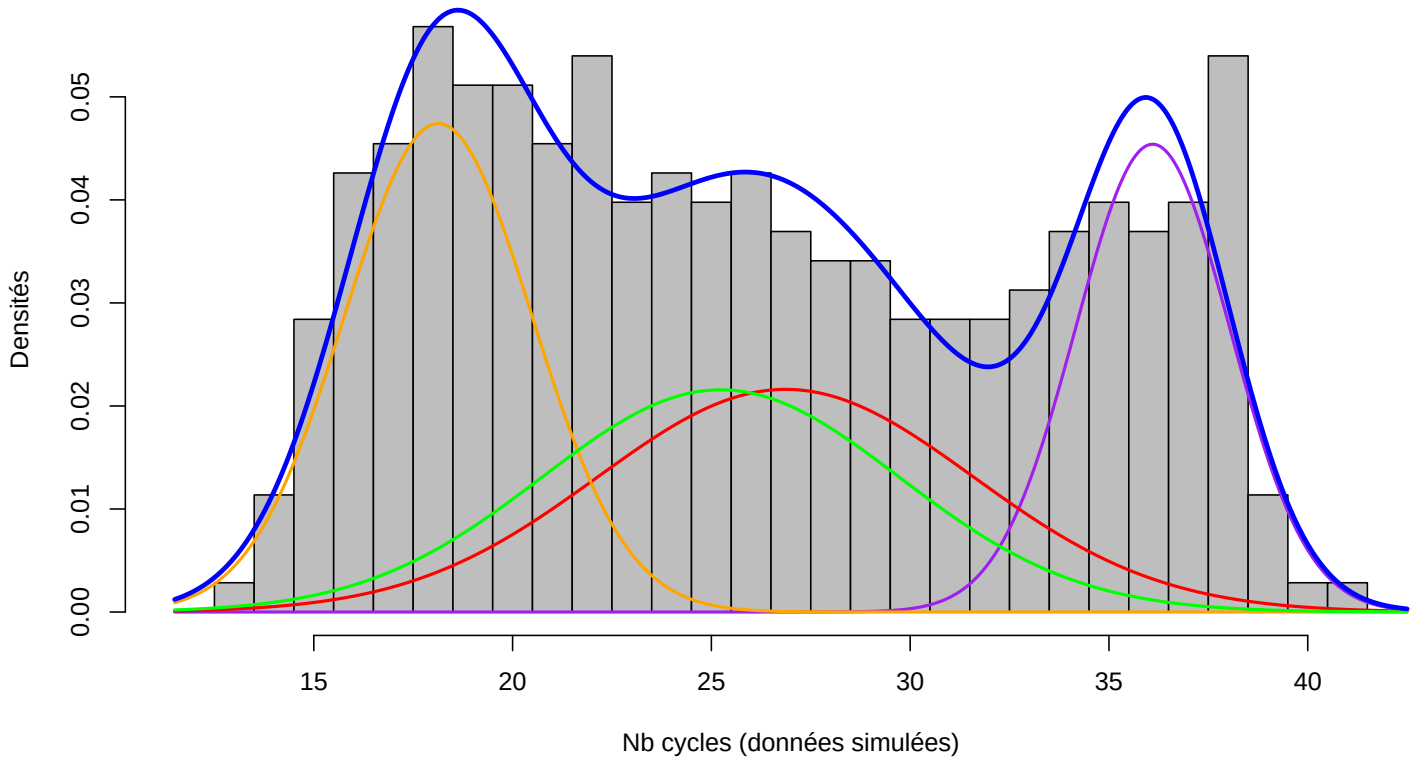


Figure: Reconstitution de l'histogramme proposé par Jones et al. [2020] avec 3 712 individus. Estimation avec un mélange de quatre lois.

Quelle est la distribution du nombre de cycles ?

↪ pas du tout d'open data...

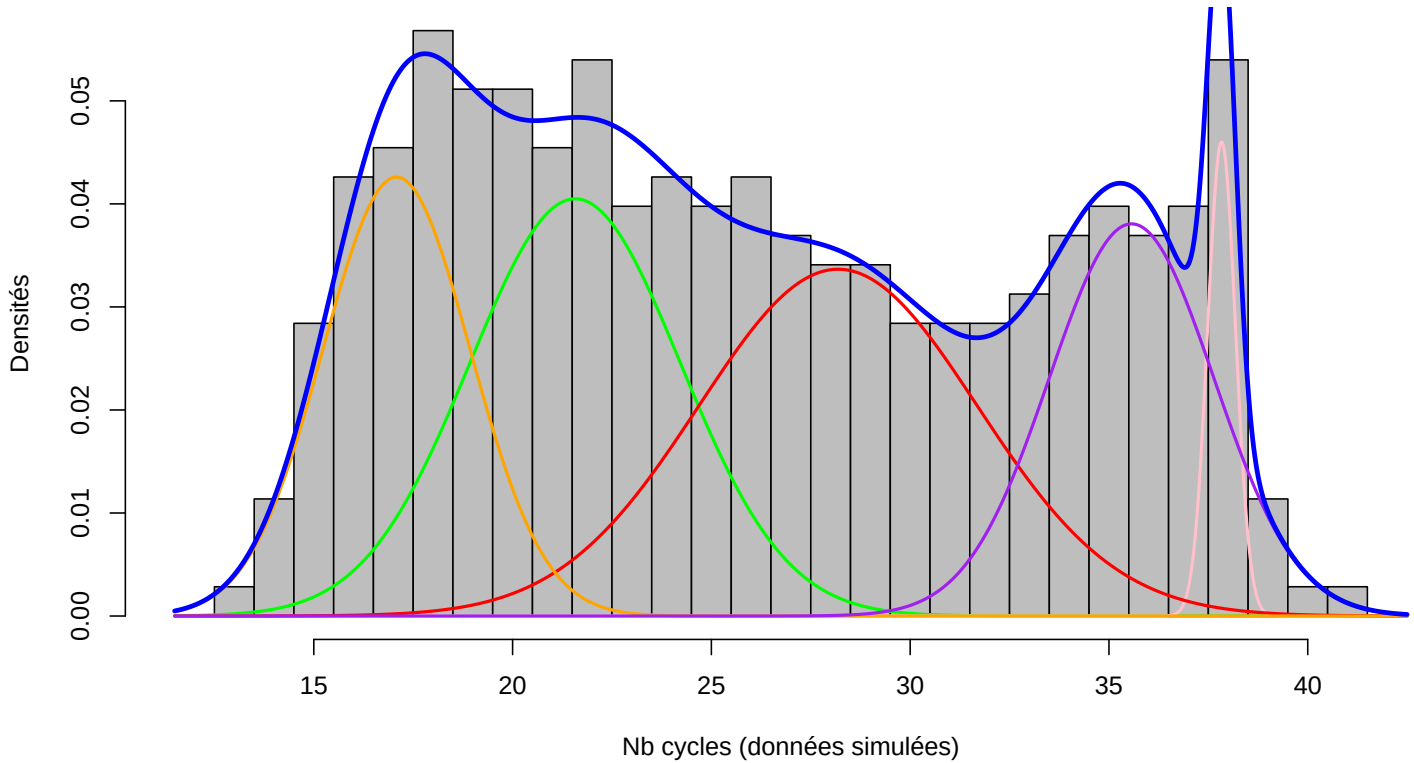


Figure: Reconstitution de l'histogramme proposé par Jones et al. [2020] avec 3 712 individus. Estimation avec un mélange de cinq lois.

Modèle de mélanges

$\hookrightarrow$ EM+BIC (Rmixmod de Lebret et al. [2015]) sur 100 simulations de jeux de données.

▸ Modèle de mélange

Modèle de mélanges

$\hookrightarrow$ EM+BIC (Rmixmod de Lebret et al. [2015]) sur 100 simulations de jeux de données.

Table: Nombre de fois que chaque modèle a été sélectionné.

K	2	3
Nb	5%	95%

Modèle de mélange

Modèle de mélanges

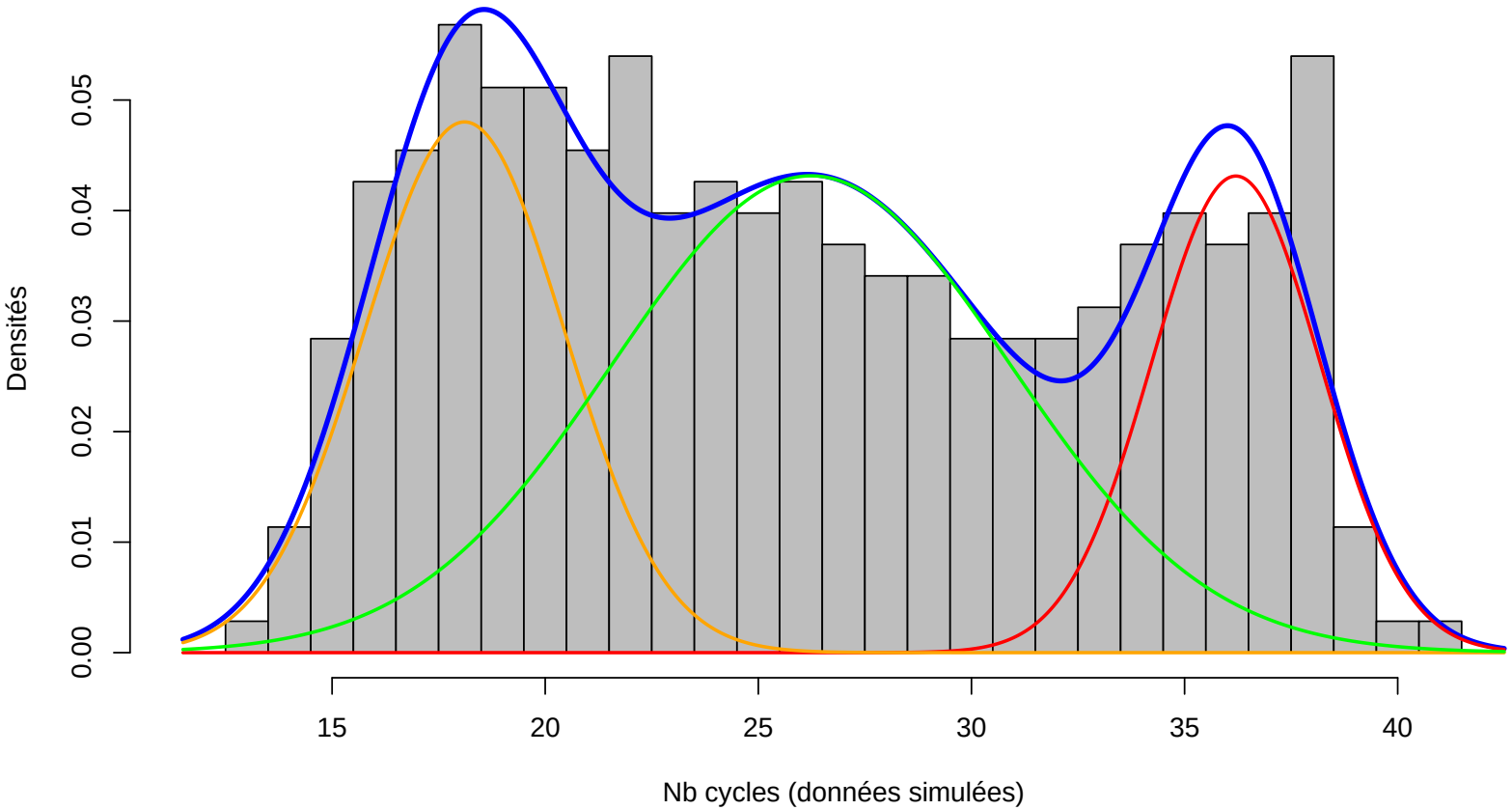


Figure: Reconstitution de l'histogramme proposé par Jones et al. [2020] avec 3 712 individus avec estimation de trois classes.

Modèle de mélange

Modèle de mélanges

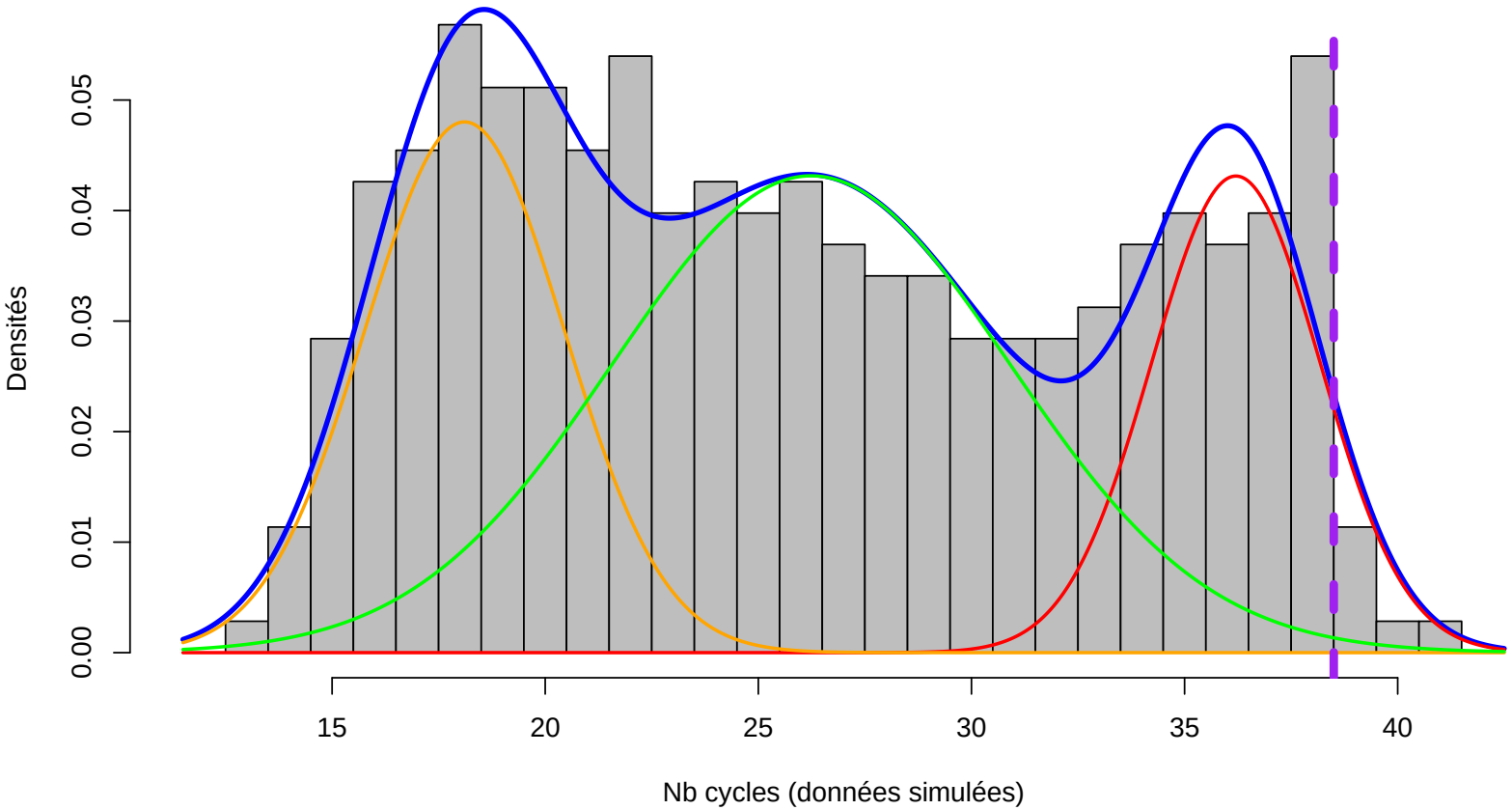
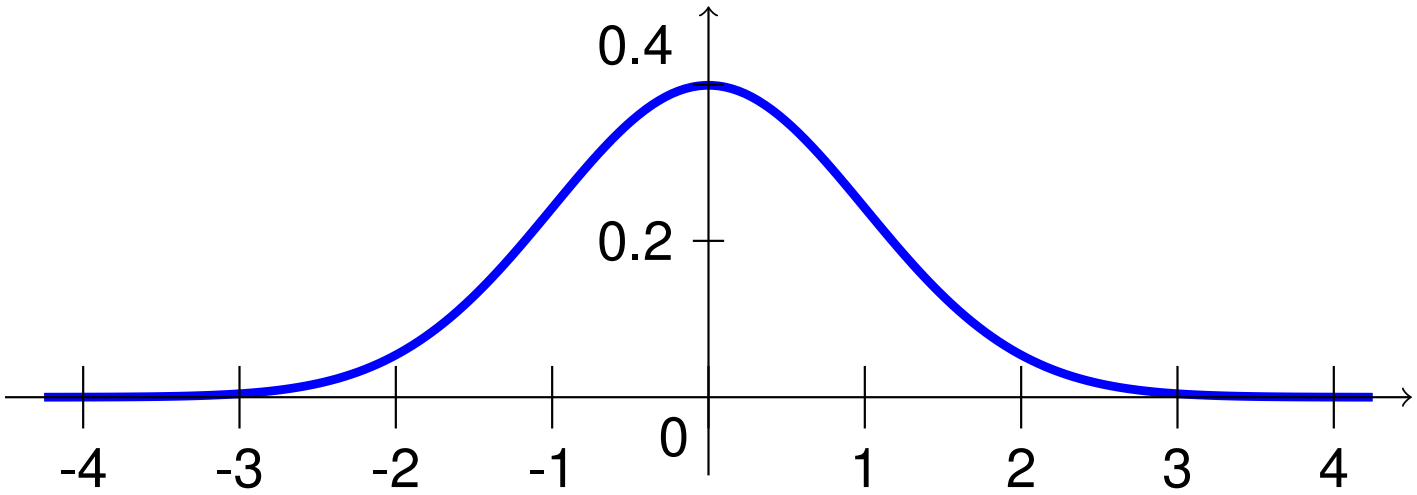


Figure: Reconstitution de l'histogramme proposé par Jones et al. [2020] avec 3 712 individus avec estimation de trois classes.

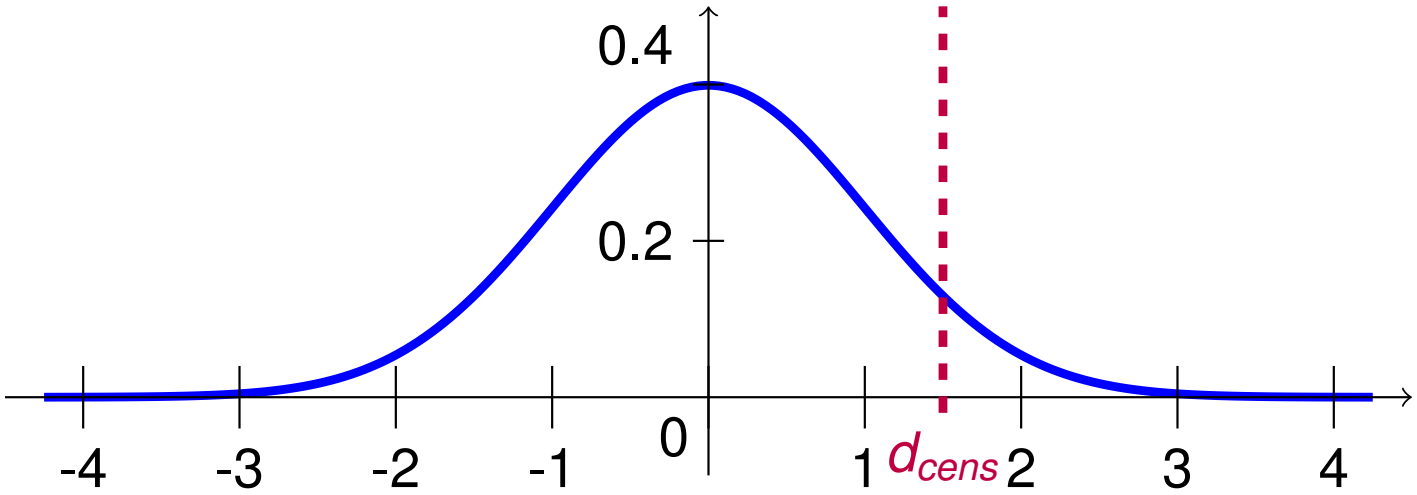
Modèle de mélange

Modèle gaussien avec censure partielle



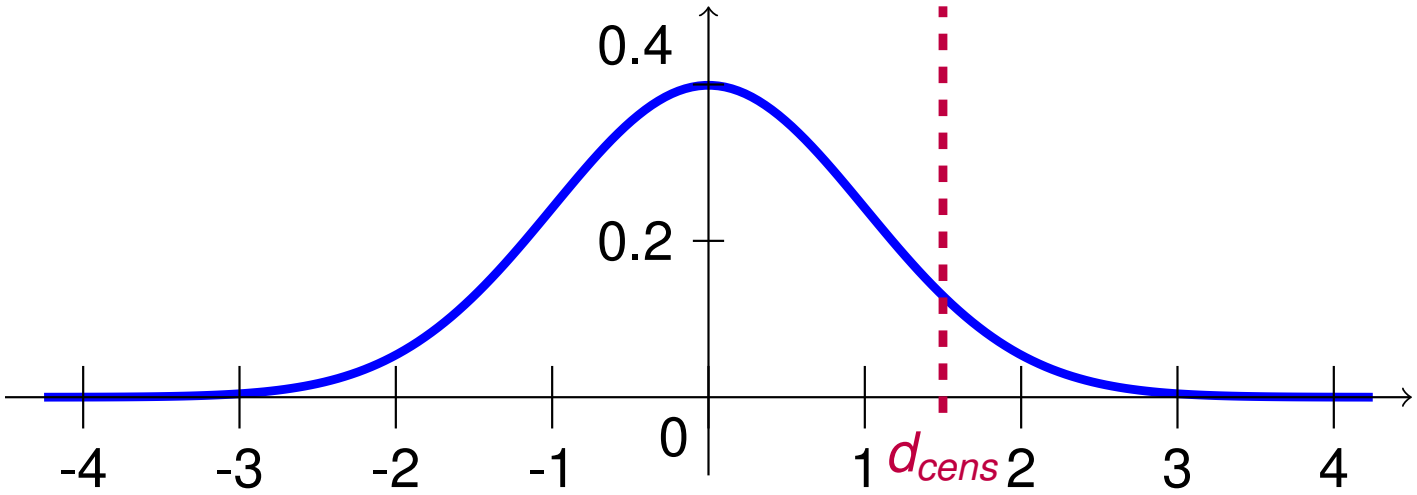
$$f_{\mu,\sigma}(x) \propto \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

Modèle gaussien avec censure partielle



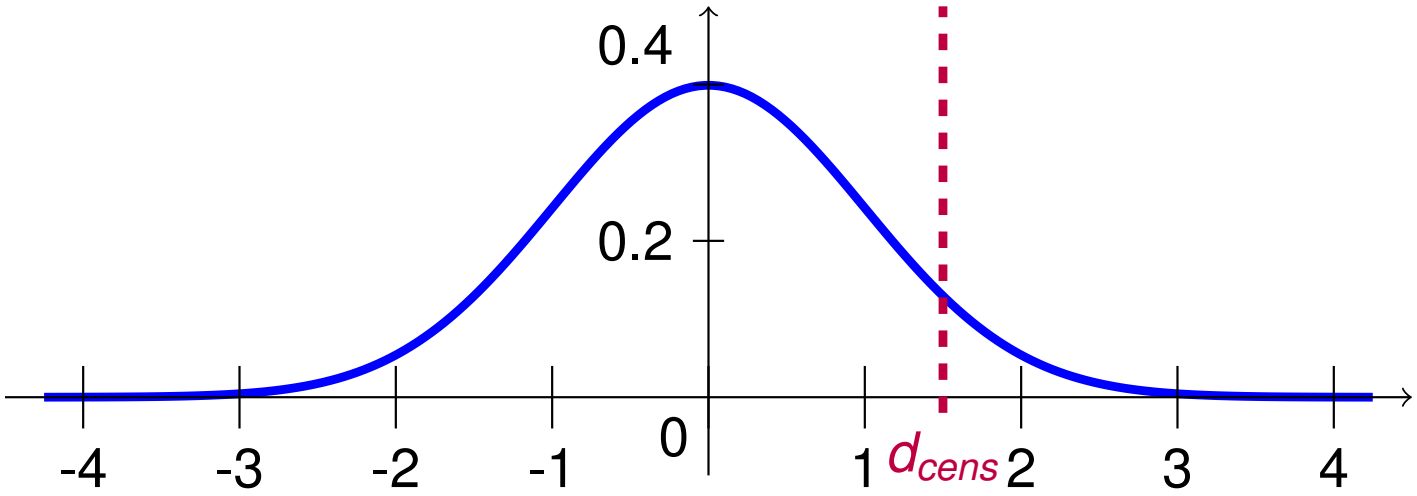
$$f_{\mu,\sigma}(x) \propto \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

Modèle gaussien avec censure partielle



$$f_{\mu,\sigma}(x) \propto \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \times \begin{cases} 1 & \text{si } x \leq d_{cens}, \end{cases}$$

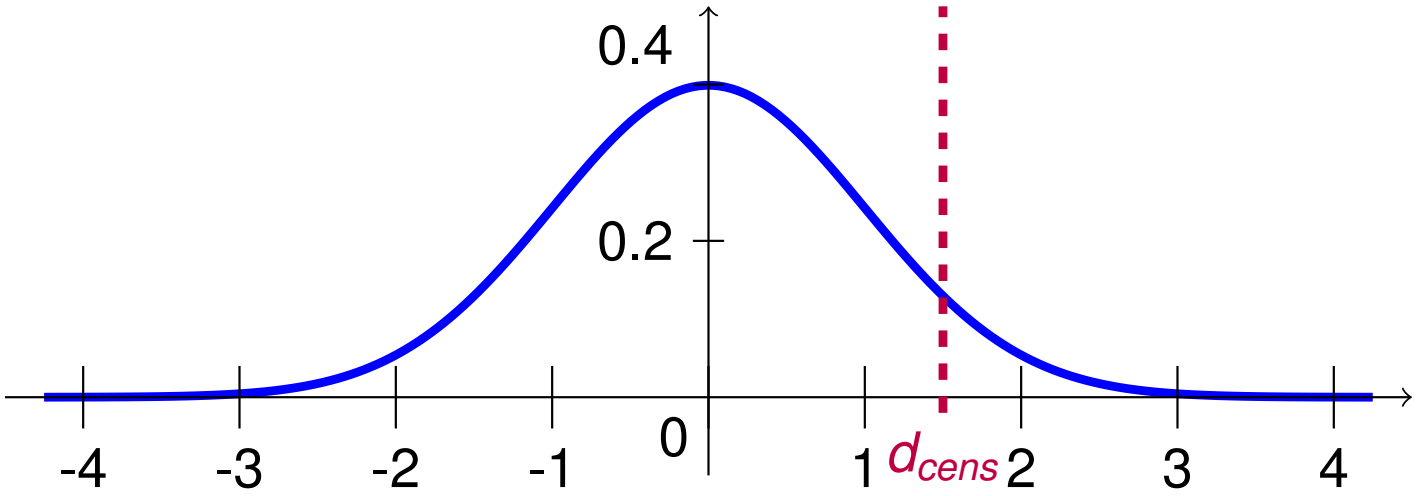
Modèle gaussien avec censure partielle



$$f_{\mu,\sigma,q}(x) \propto \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \times \begin{cases} 1 & \text{si } x \leq d_{cens}, \\ q & \text{sinon} \end{cases}$$

avec $q \in]0;1[$.

Modèle gaussien avec censure partielle



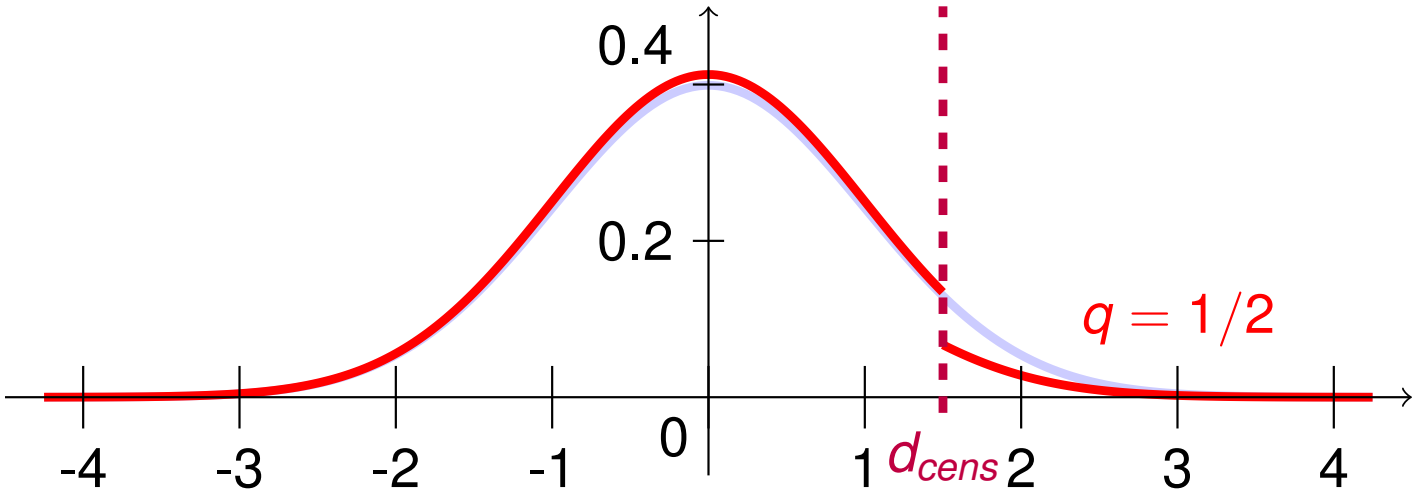
$$f_{\mu,\sigma,q}(x) \propto \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \times \begin{cases} 1 & \text{si } x \leq d_{cens}, \\ q & \text{sinon} \end{cases}$$

avec $q \in]0;1[$. Donc, nous avons :

$$f_{\mu,\sigma,q}(x) = \frac{f_{\mu,\sigma}(x)}{q + (1-q)F_{\mu,\sigma}(d_{cens})} \left[1 + (q-1)\mathbb{1}_{\{x > d_{cens}\}}(x)\right]$$

avec $f_{\mu,\sigma}$ (resp. $F_{\mu,\sigma}$) la densité (resp. la fonction de répartition) d'une loi gaussien $\mathcal{N}(\mu, \sigma^2)$.

Modèle gaussien avec censure partielle



$$f_{\mu,\sigma,q}(x) \propto \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \times \begin{cases} 1 & \text{si } x \leq d_{cens}, \\ q & \text{sinon} \end{cases}$$

avec $q \in]0;1[$. Donc, nous avons :

$$f_{\mu,\sigma,q}(x) = \frac{f_{\mu,\sigma}(x)}{q + (1-q)F_{\mu,\sigma}(d_{cens})} \left[1 + (q-1)\mathbb{1}_{\{x > d_{cens}\}}(x)\right]$$

avec $f_{\mu,\sigma}$ (resp. $F_{\mu,\sigma}$) la densité (resp. la fonction de répartition) d'une loi gaussien $\mathcal{N}(\mu, \sigma^2)$.

Modèle gaussien avec censure partielle/totale

1. Modèle gaussien avec censure partielle $\mathcal{CN}_{d_{cens}}(\mu, \sigma, q)$:

$$f_{\mu,\sigma,q}(x) = \frac{f_{\mu,\sigma}(x)}{q + (1-q)F_{\mu,\sigma}(d_{cens})}\left[1 + (q-1)\mathbb{1}_{\{x > d_{cens}\}}(x)\right]$$

2. Modèle gaussien avec censure totale $\mathcal{CN}_{d_{cens}}(\mu, \sigma) = \mathcal{CN}_{d_{cens}}(\mu, \sigma, 0)$:

$$f_{\mu,\sigma,q}(x) = \frac{f_{\mu,\sigma}(x)}{F_{\mu,\sigma}(d_{cens})}\mathbb{1}_{\{x \leq d_{cens}\}}(x)$$

Modèle gaussien avec censure partielle/totale

1. Modèle gaussien avec censure partielle $\mathcal{CN}_{d_{cens}}(\mu, \sigma, q)$:

$$f_{\mu,\sigma,q}(x) = \frac{f_{\mu,\sigma}(x)}{q + (1-q)F_{\mu,\sigma}(d_{cens})}\left[1 + (q-1)\mathbb{1}_{\{x > d_{cens}\}}(x)\right]$$

2. Modèle gaussien avec censure totale $\mathcal{CN}_{d_{cens}}(\mu, \sigma) = \mathcal{CN}_{d_{cens}}(\mu, \sigma, 0)$:

$$f_{\mu,\sigma,q}(x) = \frac{f_{\mu,\sigma}(x)}{F_{\mu,\sigma}(d_{cens})}\mathbb{1}_{\{x \leq d_{cens}\}}(x)$$

3. Un autre type ? Décroissance en 1/x ? Exponentielle ? Probabilités q décroissantes par palier pour différentes machines ?

Propriétés

Identifiabilité

Le seuil de censure d_{cens} étant fixé, les modèles gaussiens avec censure partielle et totale sont idenfitifiable.

▸ Base des démonstrations

Propriétés

Identifiabilité

Le seuil de censure d_{cens} étant fixé, les modèles gaussiens avec censure partielle et totale sont idenfitifiable.

Estimateur du maximum de vraisemblance

Le seuil de censure d_{cens} étant fixé, l'estimateur $(\widehat{\mu}, \widehat{\sigma}, \widehat{q})$ des paramètres $(\mu^\star, \sigma^\star, q^\star)$ obtenu par maximum de vraisemblance est fortement consistant et asymptotiquement normal.

▸ Base des démonstrations

Propriétés

Identifiabilité

Le seuil de censure d_{cens} étant fixé, les modèles gaussiens avec censure partielle et totale sont idenfitifiable.

Estimateur du maximum de vraisemblance

Le seuil de censure d_{cens} étant fixé, l'estimateur $(\widehat{\mu}, \widehat{\sigma}, \widehat{q})$ des paramètres $(\mu^\star, \sigma^\star, q^\star)$ obtenu par maximum de vraisemblance est fortement consistant et asymptotiquement normal.

$\hookrightarrow$ Par contre, il n'existe pas de forme analytique.

▸ Base des démonstrations

Propriétés

Identifiabilité

Le seuil de censure d_{cens} étant fixé, les modèles gaussiens avec censure partielle et totale sont idenfitifiable.

Estimateur du maximum de vraisemblance

Le seuil de censure d_{cens} étant fixé, l'estimateur $(\widehat{\mu}, \widehat{\sigma}, \widehat{q})$ des paramètres $(\mu^\star, \sigma^\star, q^\star)$ obtenu par maximum de vraisemblance est fortement consistant et asymptotiquement normal.

$\hookrightarrow$ Par contre, il n'existe pas de forme analytique.
$\Rightarrow$ Maximisation à l'aide d'un algorithme type Newton-Raphson implémenté dans la fonction `nlm` de `R`.

▸ Base des démonstrations

Mélange de gaussiennes (partiellement) censurées

Hypothèses :

- Le seuil d_{cens} est connu et le même pour toutes les clusters.

Mélange de gaussiennes (partiellement) censurées

Hypothèses :

- Le seuil d_{cens} est connu et le même pour toutes les clusters.
- q ou q_k ?
  - Même q pour tous les clusters : logique mais plus compliqué à implémenter.

Mélange de gaussiennes (partiellement) censurées

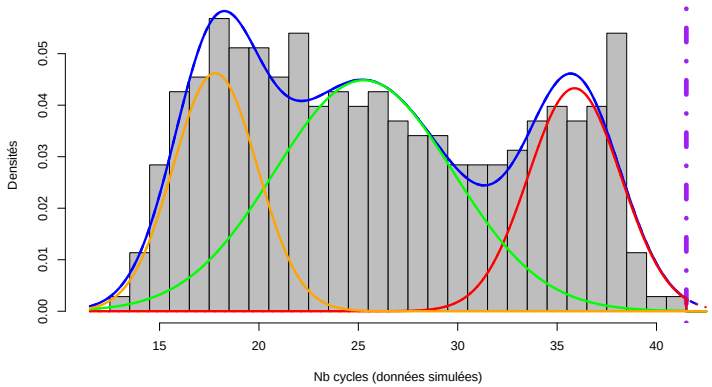
Hypothèses :

- Le seuil d_{cens} est connu et le même pour toutes les clusters.
- q ou q_k ?
  - Même q pour tous les clusters : logique mais plus compliqué à implémenter.
  - q_k dépendant du cluster k : moins cohérent mais plus facile à implémenter et parallélisable.

Plan

Application : données de Jones et al. [2020]

$d_{cens} = 41.5$



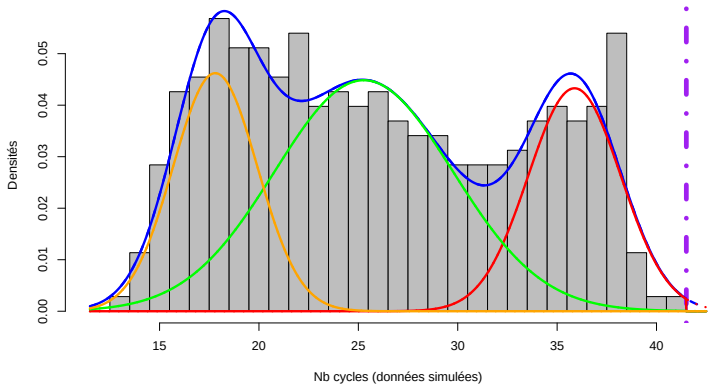
$d_{cens} = 40.5$

$d_{cens} = 39.5$

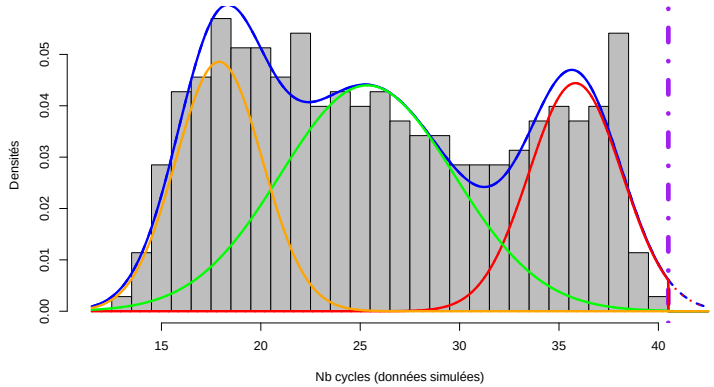
$d_{cens} = 38.5$

Application : données de Jones et al. [2020]

$d_{cens} = 41.5$



$d_{cens} = 40.5$

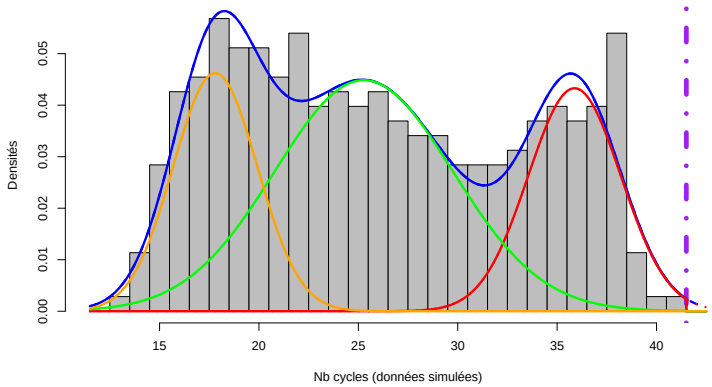


$d_{cens} = 39.5$

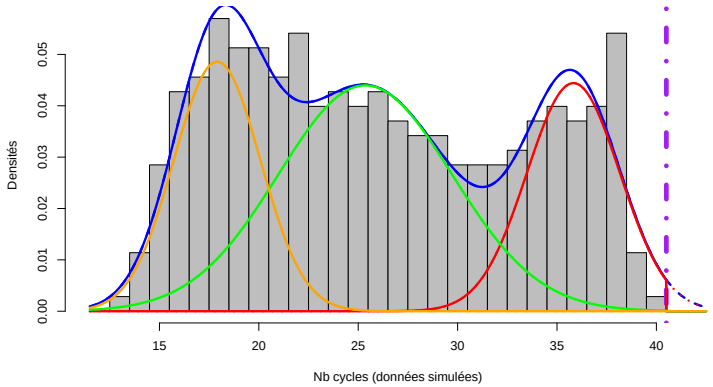
$d_{cens} = 38.5$

Application : données de Jones et al. [2020]

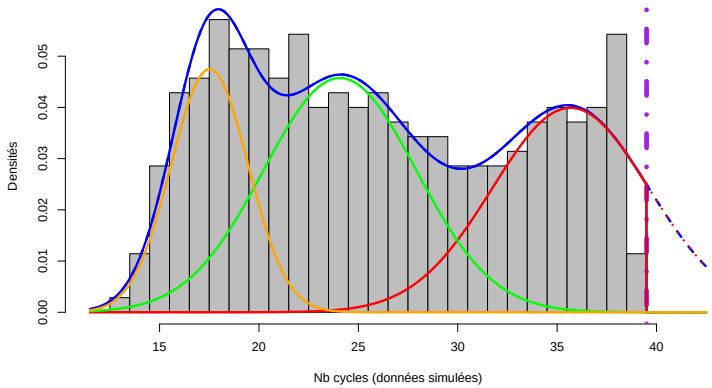
$d_{cens} = 41.5$



$d_{cens} = 40.5$



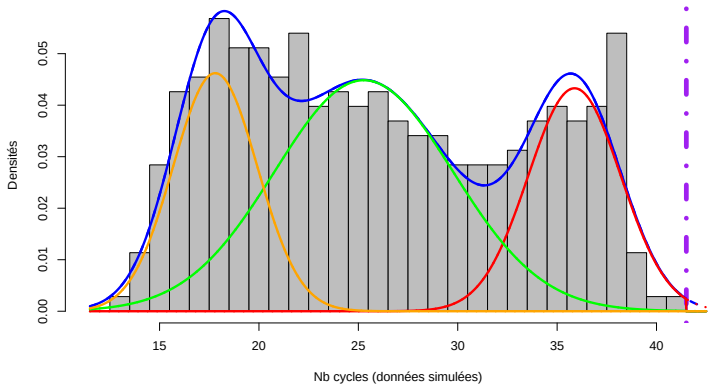
$d_{cens} = 39.5$



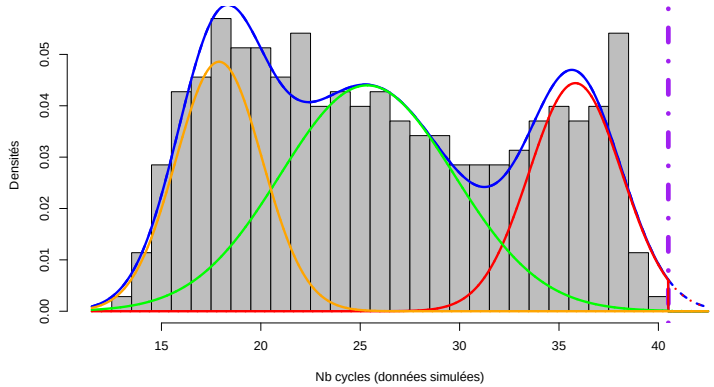
$d_{cens} = 38.5$

Application : données de Jones et al. [2020]

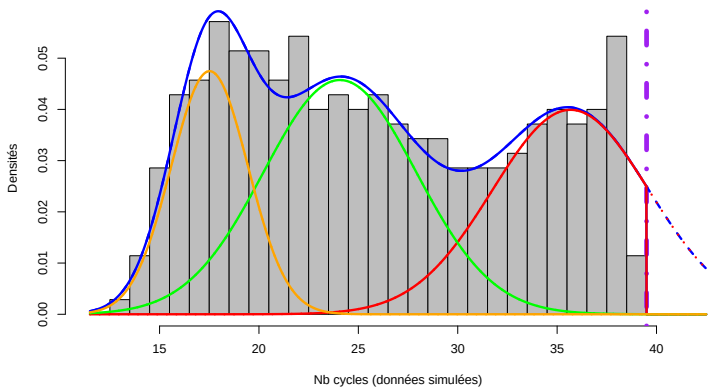
$d_{cens} = 41.5$



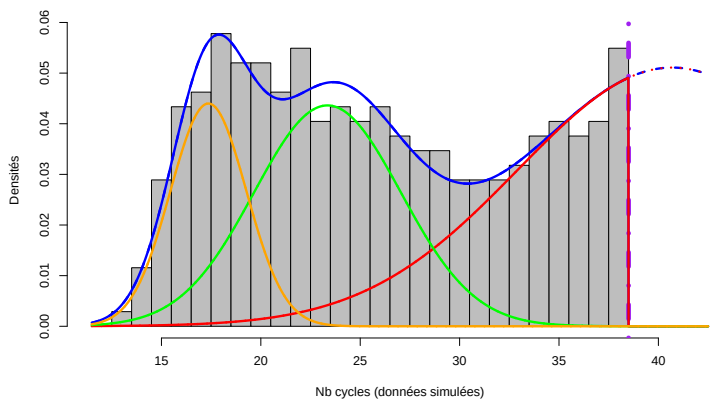
$d_{cens} = 40.5$



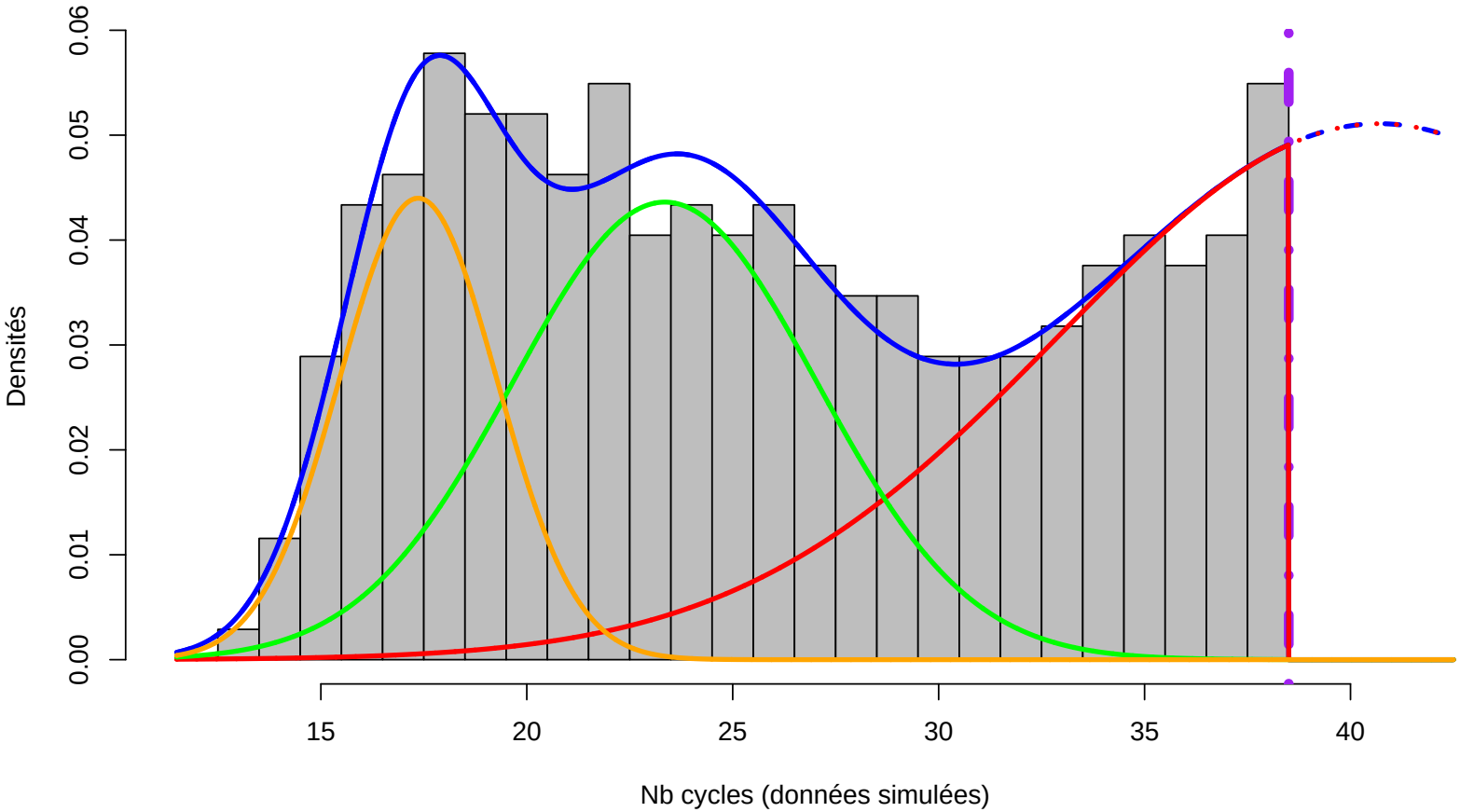
$d_{cens} = 39.5$



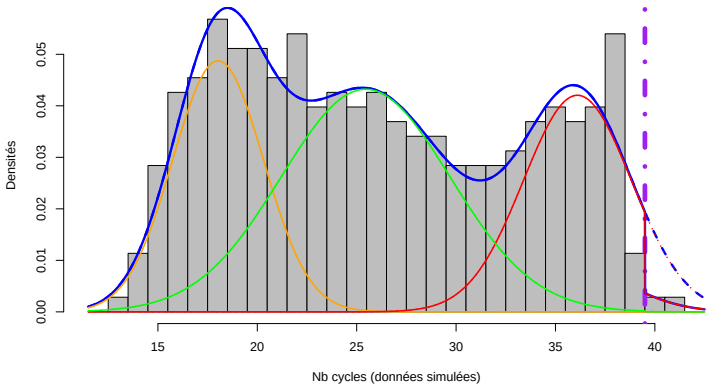
$d_{cens} = 38.5$



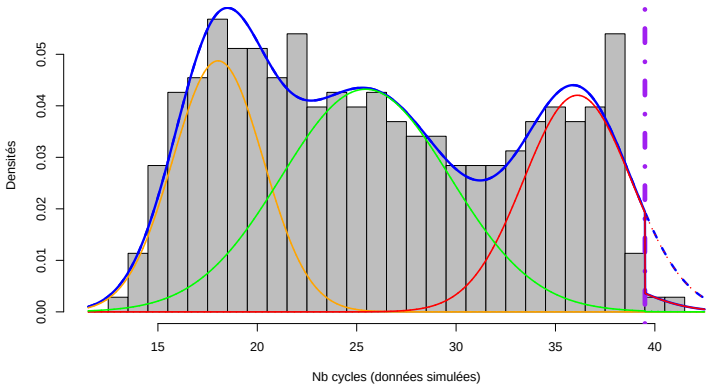
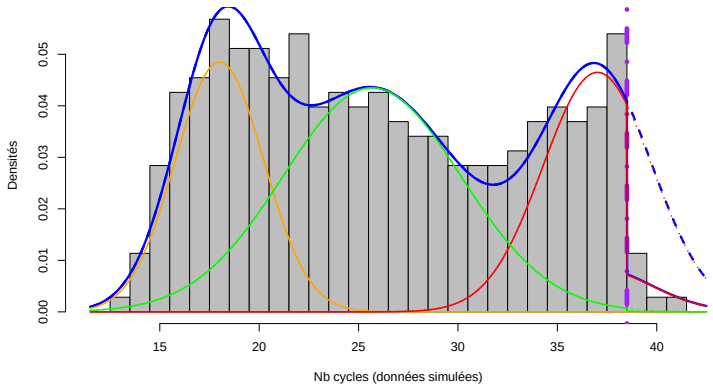
Application : données de Jones et al. [2020]



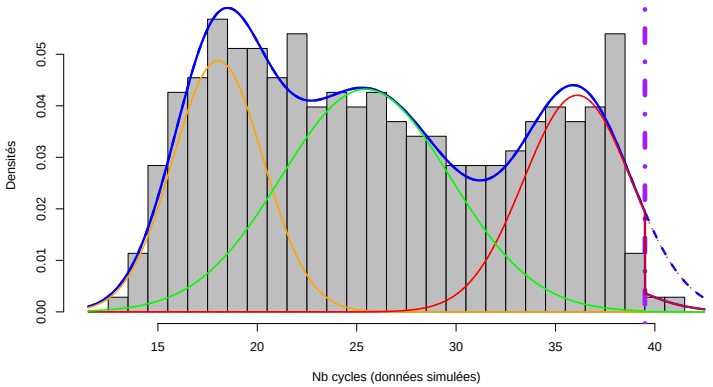
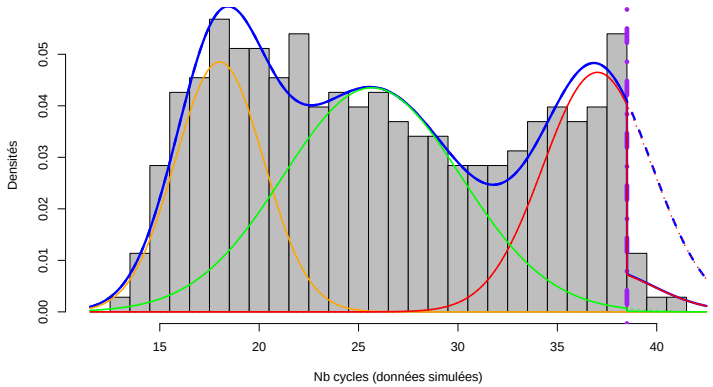
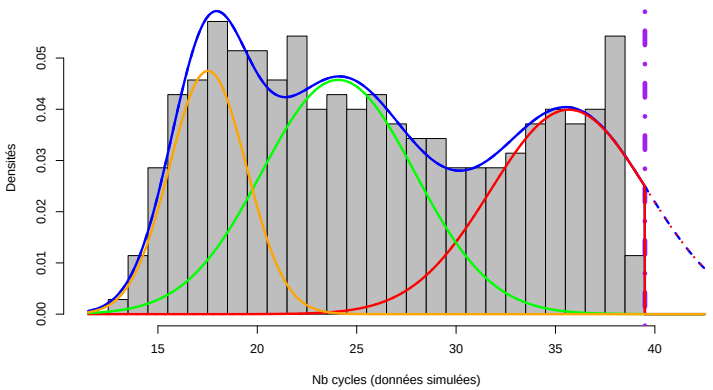
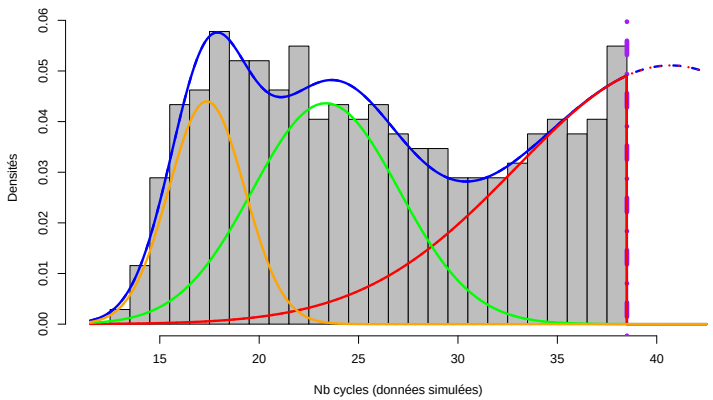
Application : données de Jones et al. [2020]

	$d_{cens} = 39.5$	$d_{cens} = 38.5$
Partielle		

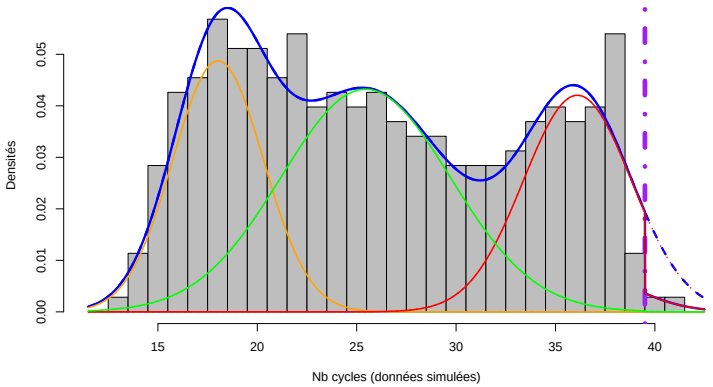
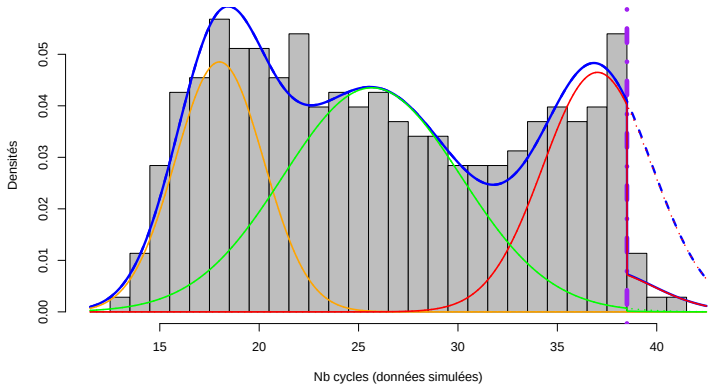
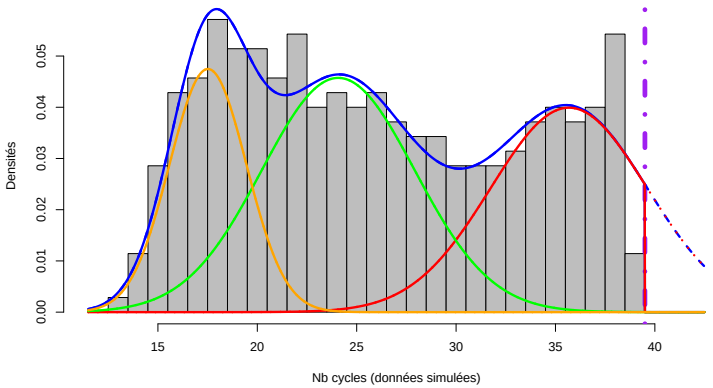
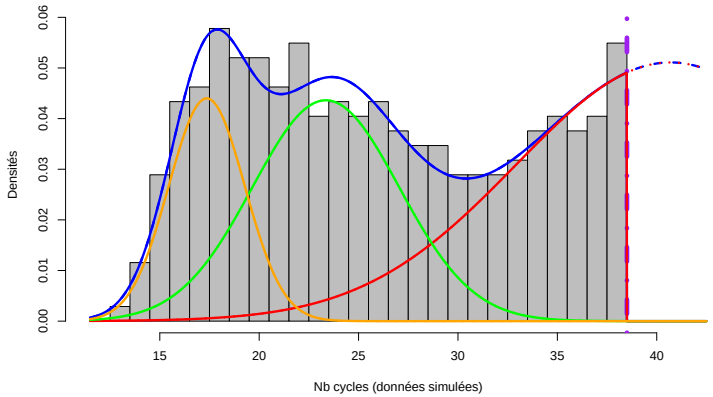
Application : données de Jones et al. [2020]

	$d_{cens} = 39.5$	$d_{cens} = 38.5$
Partielle		

Application : données de Jones et al. [2020]

	$d_{cens} = 39.5$	$d_{cens} = 38.5$
Partielle		
Complète		

Application : données de Jones et al. [2020]

	$d_{cens} = 39.5$	$d_{cens} = 38.5$
Partielle		
	Faux Neg $\geq 2.62\%$	*Faux Neg* $\geq 4.20\%$
Complète		
	Faux Neg $\geq 4.51\%$	*Faux Neg* $\geq 23.29\%$

▶ Formule

Pooling et faux négatifs

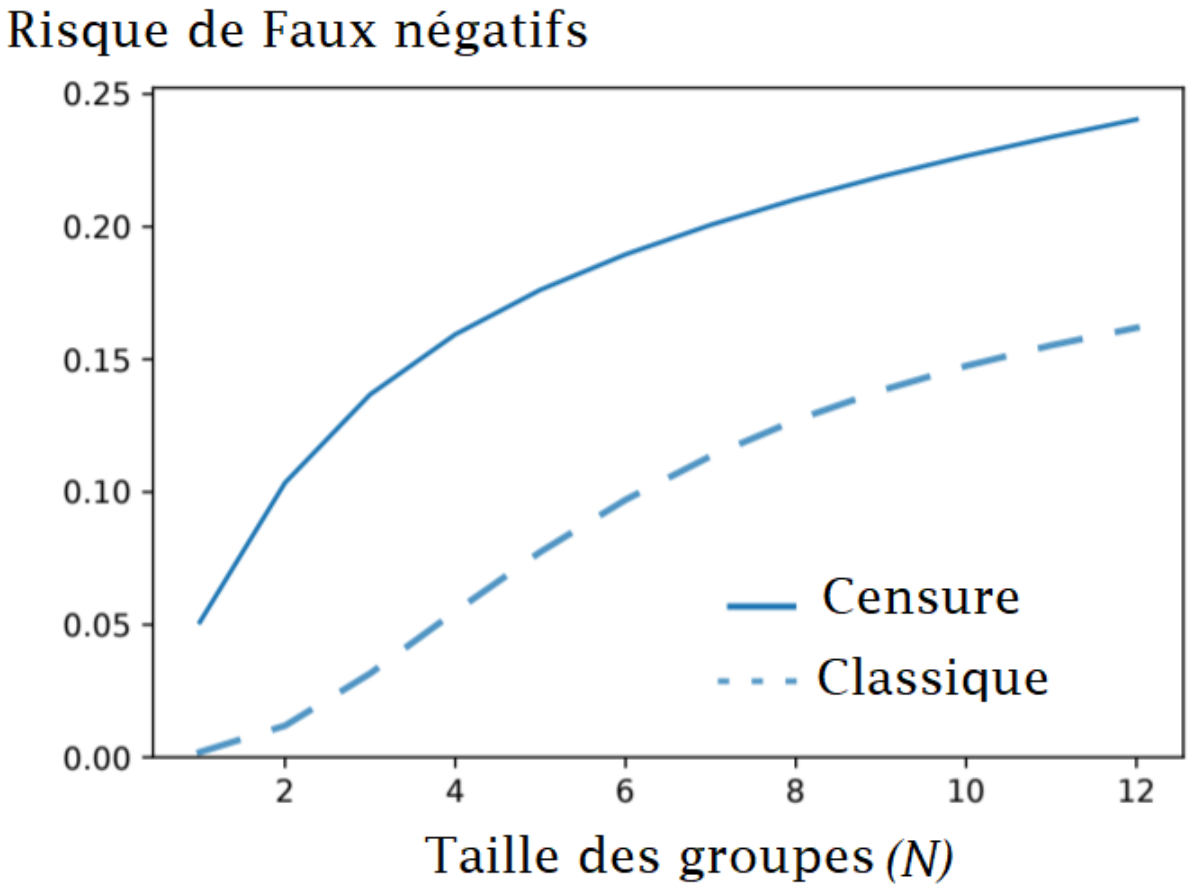


Figure: Évolution du risque de faux négatifs en fonction de la taille des groupes (abscisse) et les modélisations : modèle de mélange classique (trait pointillé) et modèle de mélange avec censure (trait plein).

Plan

Conclusions et perspectives

- Le pooling peut être utilisé pour estimer la prévalence.
- Cela pourrait aider la détection précoce de nouveaux cas dans les communautés closes.

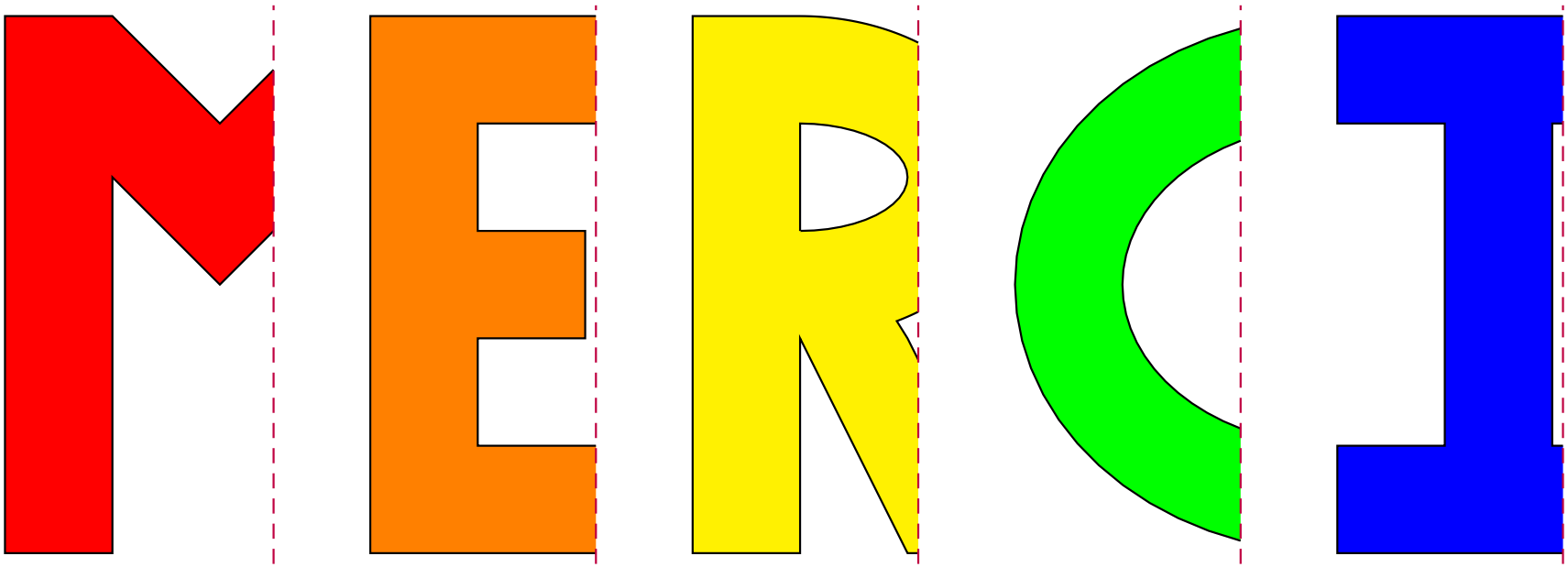
Conclusions et perspectives

- Le pooling peut être utilisé pour estimer la prévalence.
- Cela pourrait aider la détection précoce de nouveaux cas dans les communautés closes.

Perspectives :

- Amélioration de la procédure d'estimation.
- Réflexion sur les faux négatifs.
- Interprétation des clusters obtenus sur les données.

Merci pour votre attention



Bibliography I

R. Ben-Ami, A. Klochendler, M. Seidel, T. Sido, O. Gurel-Gurevich, M. Yassour, E. Meshorer, G. Benedek, I. Fogel, E. Oiknine-Djian, et al. Large-scale implementation of pooled rna extraction and rt-pcr for sars-cov-2 detection. Clinical Microbiology and Infection, 26(9):1248–1253, 2020.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1):1–22, 1977.

R. Dorfman. The detection of defective members of large populations. The Annals of Mathematical Statistics, 14(4):436–440, 1943.

T. C. Jones, B. Mühlemann, T. Veith, G. Biele, M. Zuchowski, J. Hoffmann, A. Stein, A. Edelmann, V. M. Corman, and C. Drosten. An analysis of sars-cov-2 viral load by patient age. medRxiv, 2020.

R. Lebret, S. Iovleff, F. Langrognet, C. Biernacki, G. Celeux, and G. Govaert. Rmixmod: the r package of the model-based unsupervised, supervised and semi-supervised classification mixmod library. 2015.

Bibliography II

L. Mutesa, P. Ndishimye, Y. Butera, J. Souopgui, A. Uwineza, R. Rutayisire, E. L. Ndoricimpaye, E. Musoni, N. Rujeni, T. Nyatanyi, et al. A pooled testing strategy for identifying sars-cov-2 at low prevalence. Nature, pages 1–5, 2020.

G. Schwarz et al. Estimating the dimension of a model. The annals of statistics, 6(2):461–464, 1978.

R. C. Team. R core team. r: A language and environment for statistical computing. Foundation for Statistical Computing, 2013.

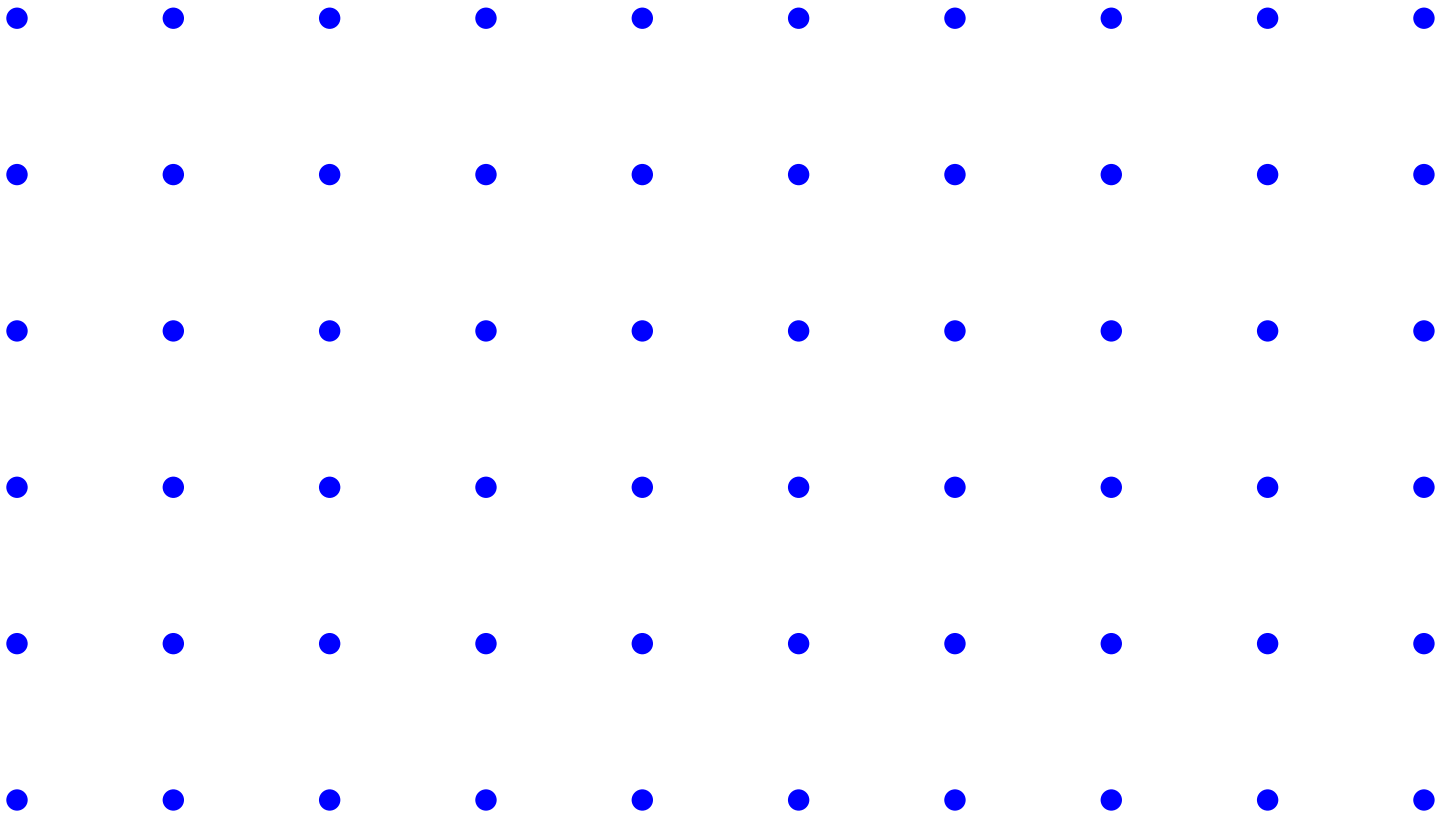
K. H. Thompson. Estimation of the proportion of vectors in a natural population of insects. Biometrics, 18(4):568–578, 1962.

I. Yelin, N. Aharony, E. Shaer-Tamar, A. Argoetti, E. Messer, D. Berenbaum, E. Shafran, A. Kuzli, N. Gandali, T. Hashimshony, et al. Evaluation of covid-19 rt-qpcr test in multi-sample pools. MedRxiv, 2020.

Plan

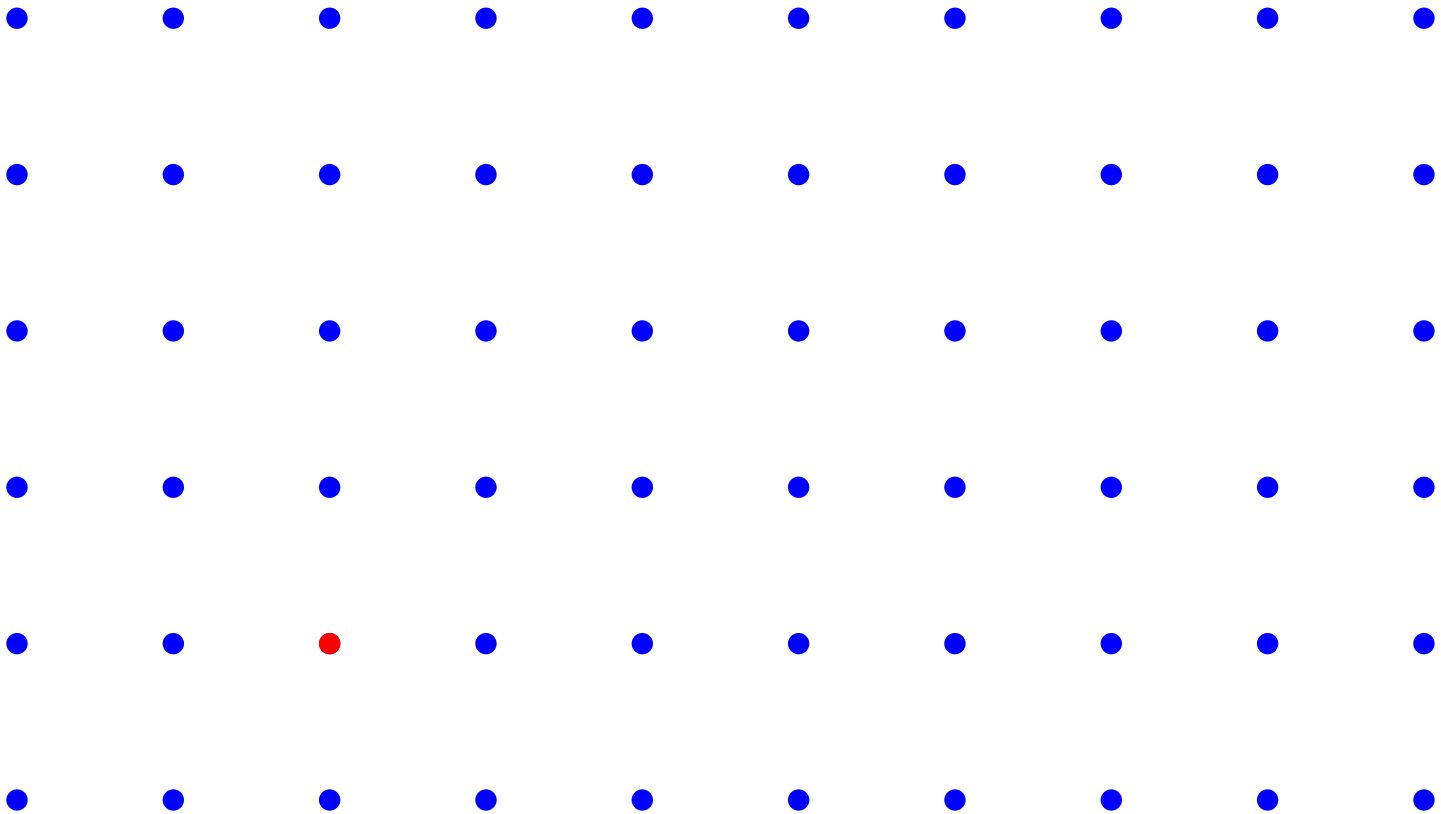
Test parfait : quelle stratégie ?

Prévalence de $p \in]0;1[$, individu $X_i \sim \mathcal{B}(p)$ d'être malade/positif.



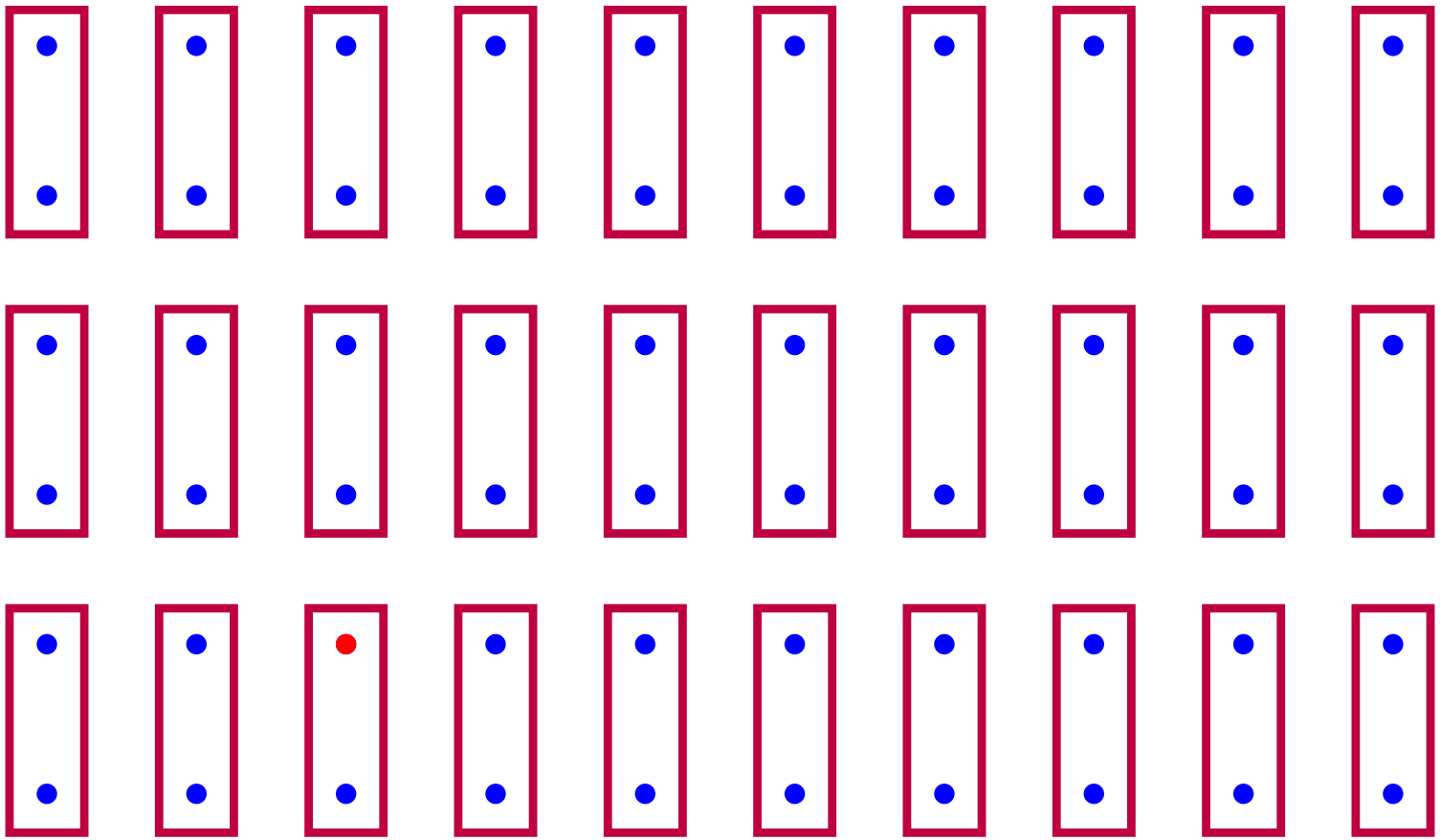
Test parfait : quelle stratégie ?

Prévalence de $p \in]0;1[$, individu $X_i \sim \mathcal{B}(p)$ d'être malade/positif.



Test parfait : quelle stratégie ?

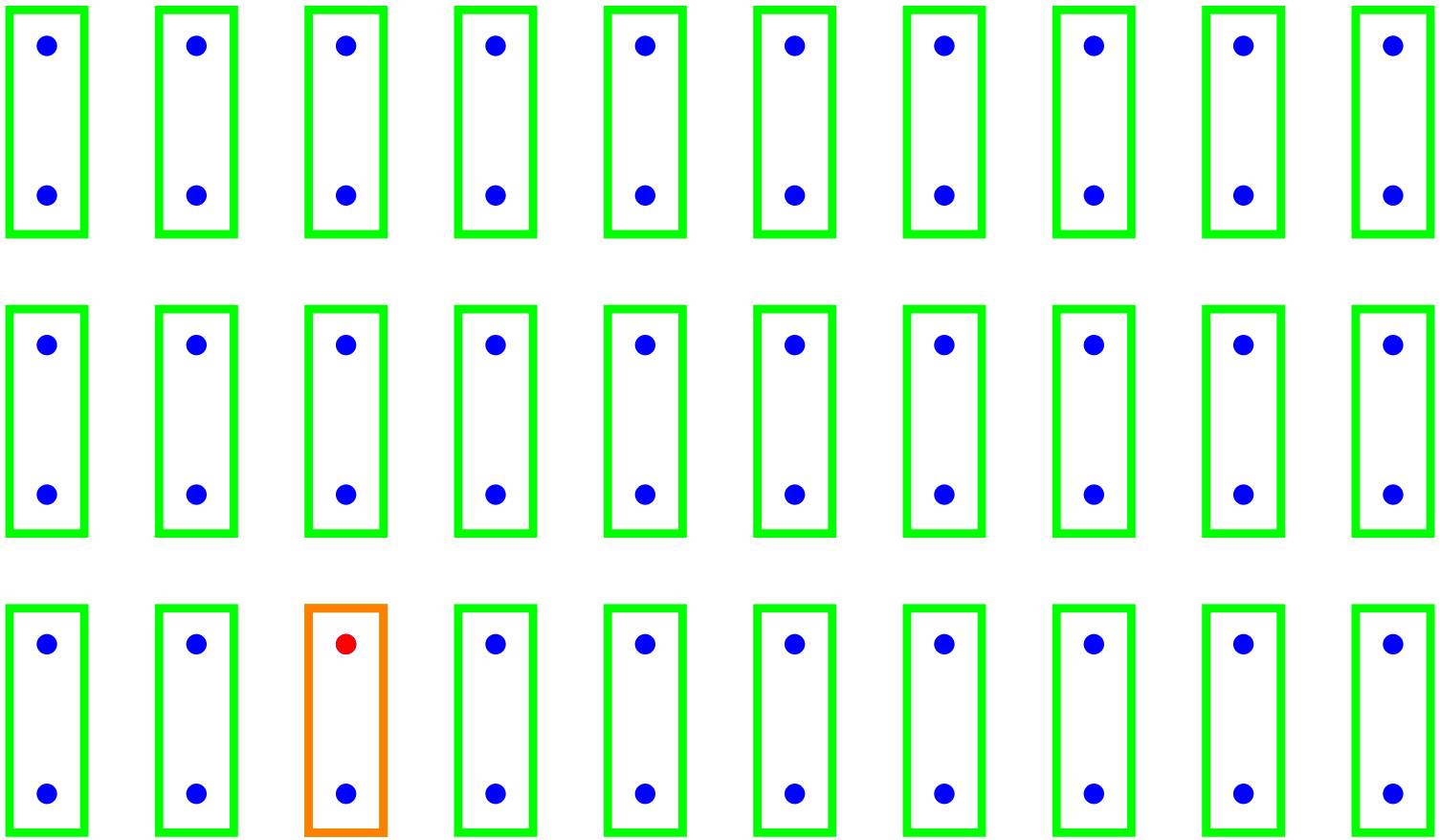
Prévalence de $p \in]0;1[$, individu $X_i \sim \mathcal{B}(p)$ d'être malade/positif.



30 tests groupés +

Test parfait : quelle stratégie ?

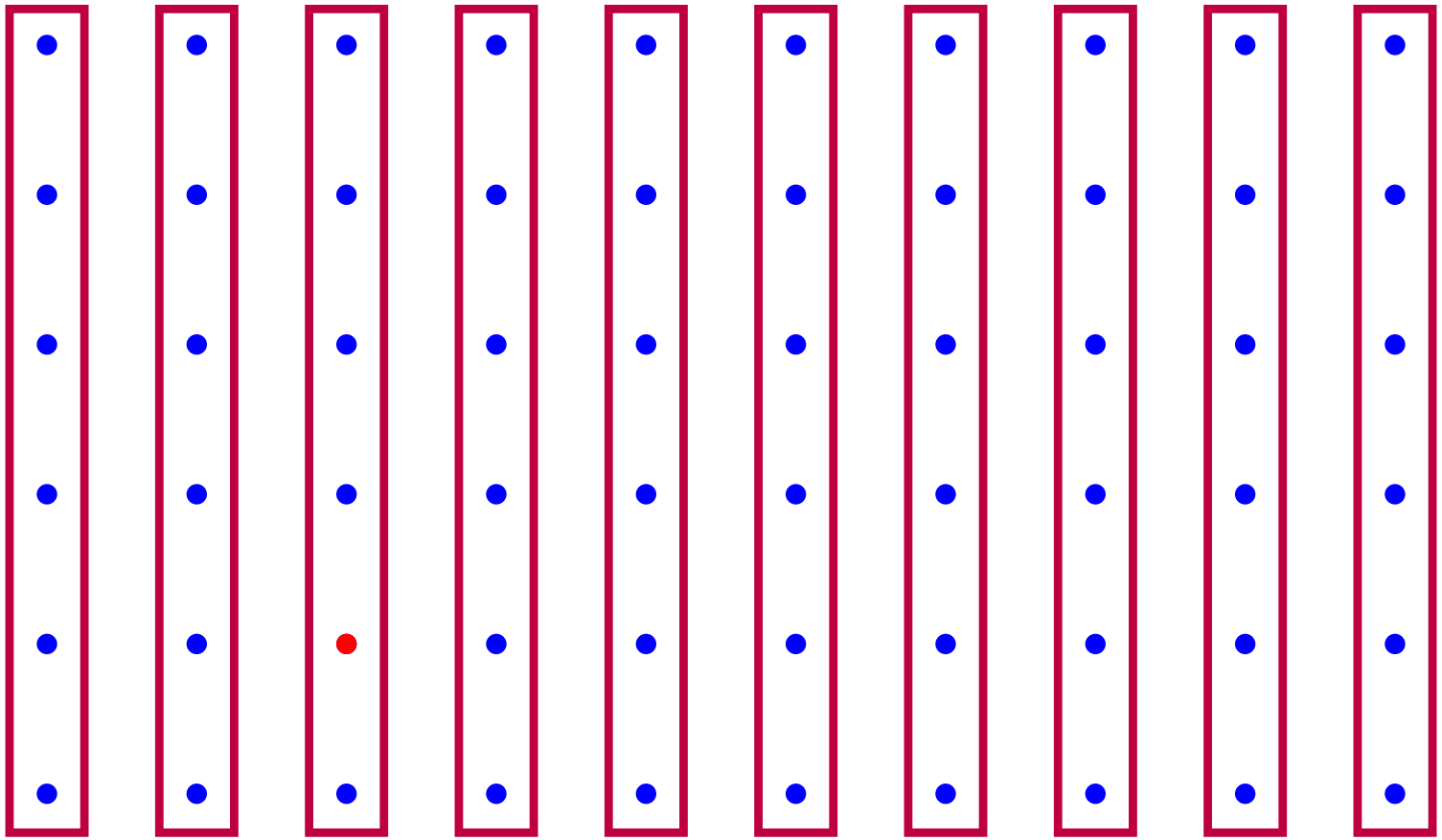
Prévalence de $p \in]0;1[$, individu $X_i \sim \mathcal{B}(p)$ d'être malade/positif.



30 tests groupés + 2 individuels

Test parfait : quelle stratégie ?

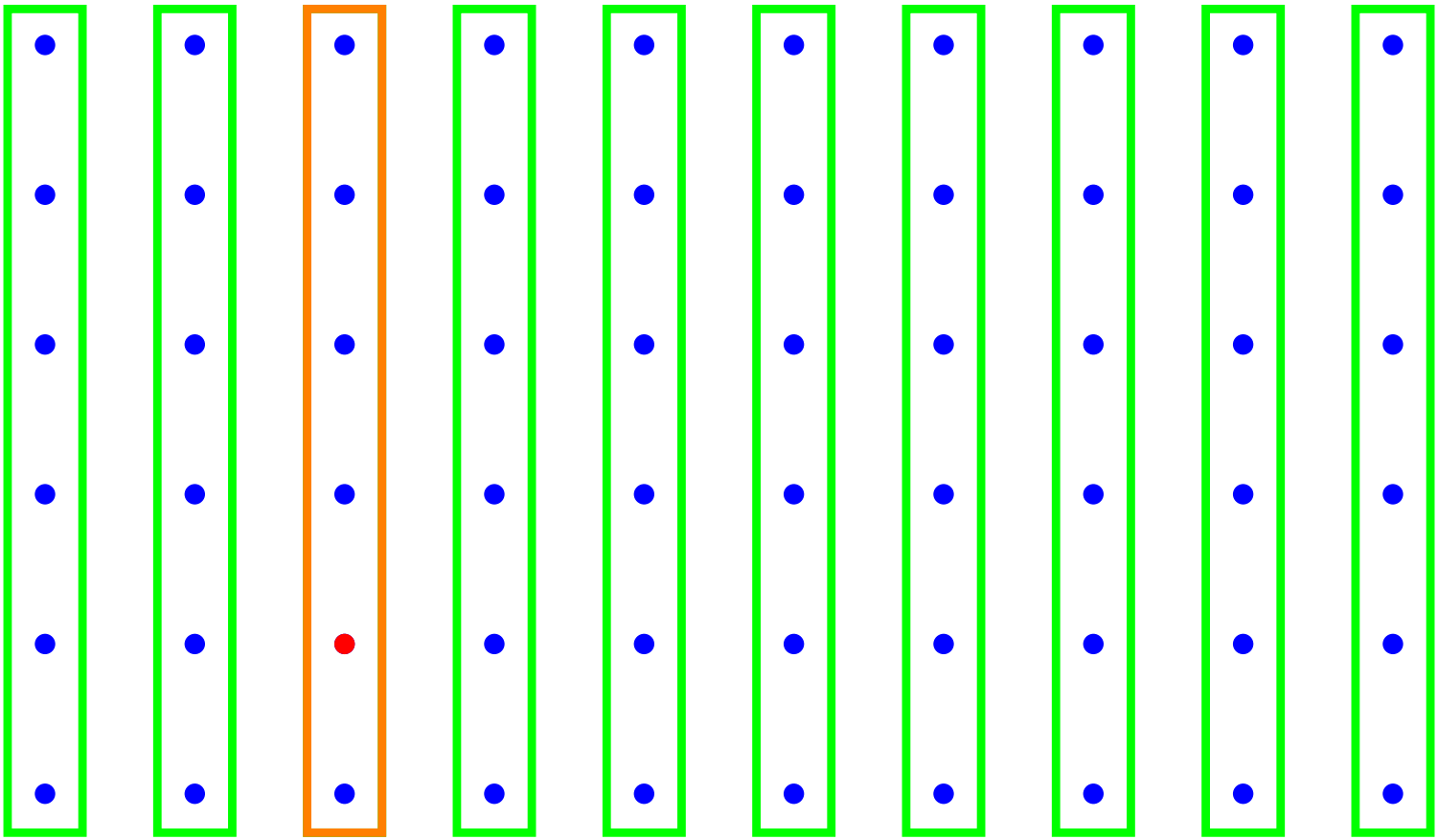
Prévalence de $p \in]0;1[$, individu $X_i \sim \mathcal{B}(p)$ d'être malade/positif.



10 tests groupés

Test parfait : quelle stratégie ?

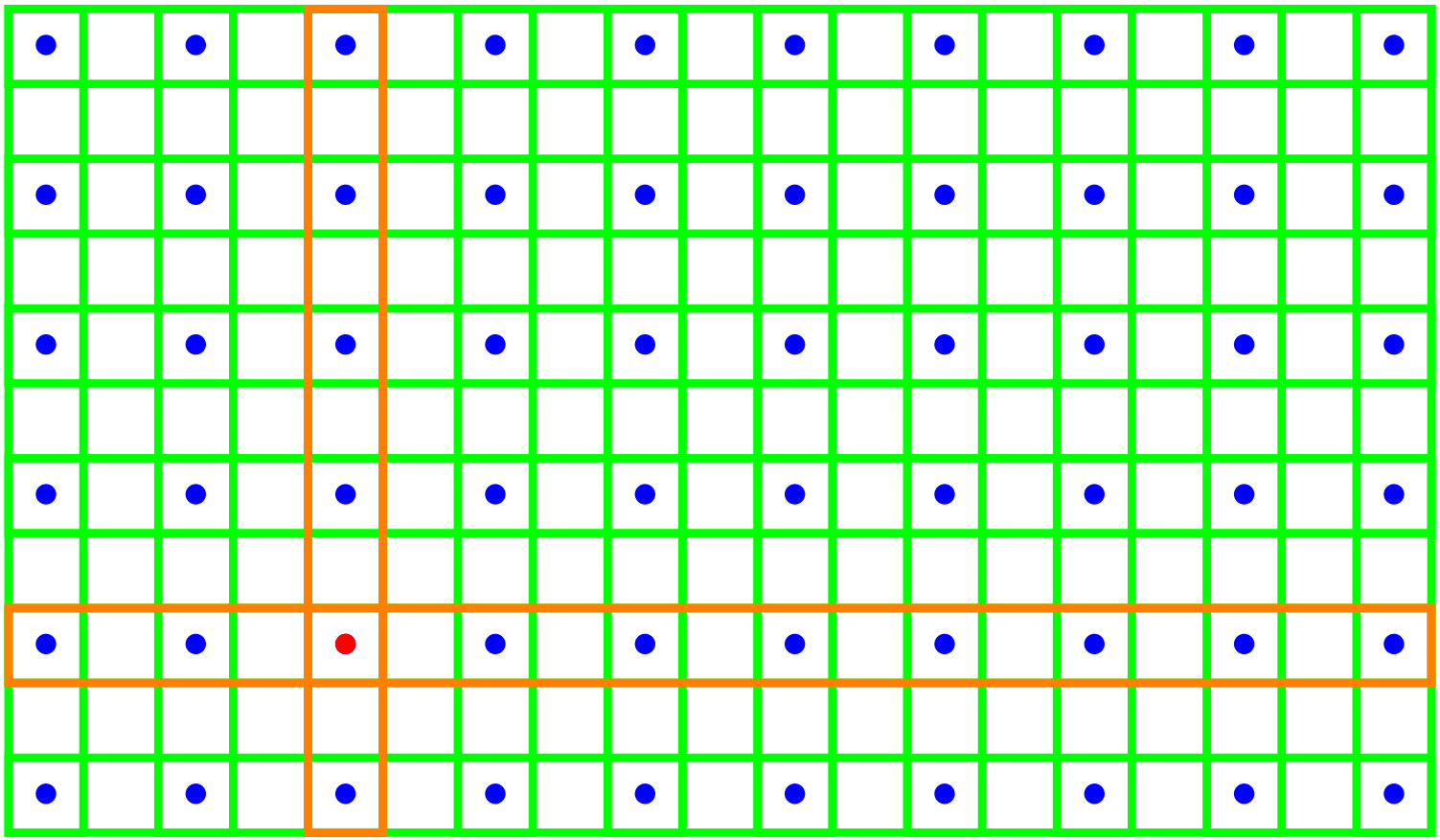
Prévalence de $p \in]0;1[$, individu $X_i \sim \mathcal{B}(p)$ d'être malade/positif.



10 tests groupés + 6 individuels

Test parfait : quelle stratégie ?

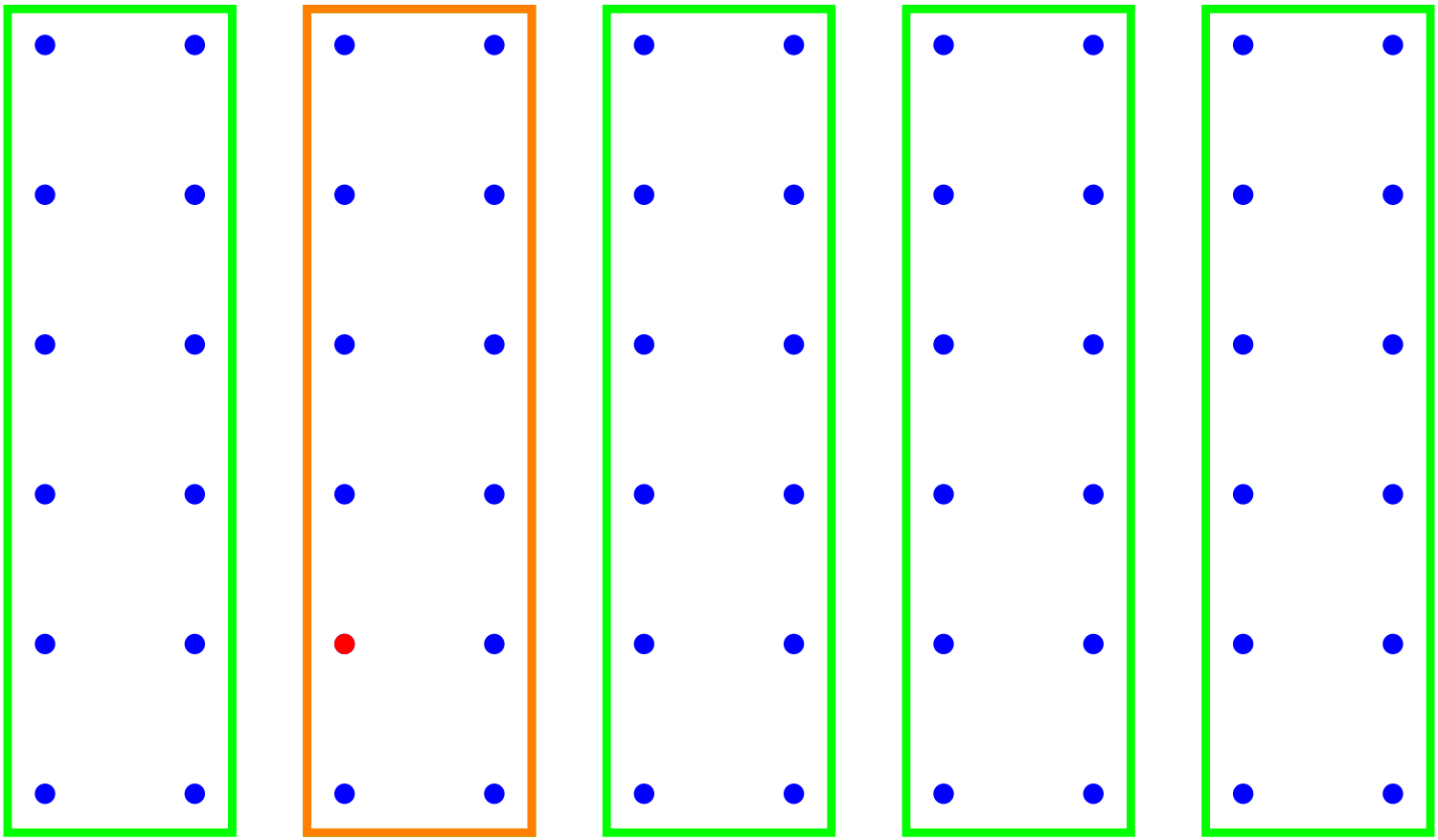
Prévalence de $p \in]0;1[$, individu $X_i \sim \mathcal{B}(p)$ d'être malade/positif.



Ou 16 tests groupés

Test parfait : quelle stratégie ?

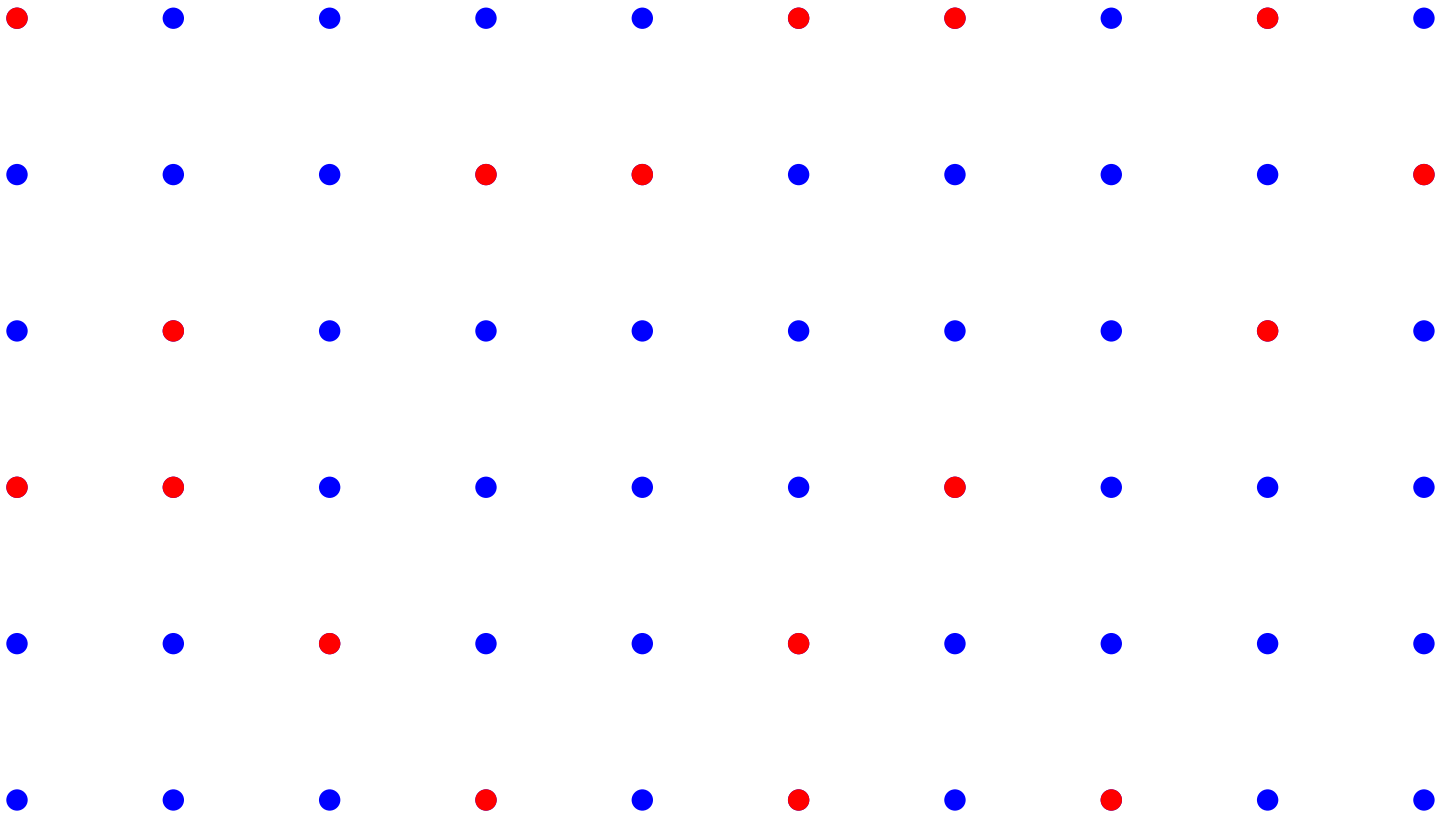
Prévalence de $p \in]0;1[$, individu $X_i \sim \mathcal{B}(p)$ d'être malade/positif.



5 tests groupés + 12 individuels

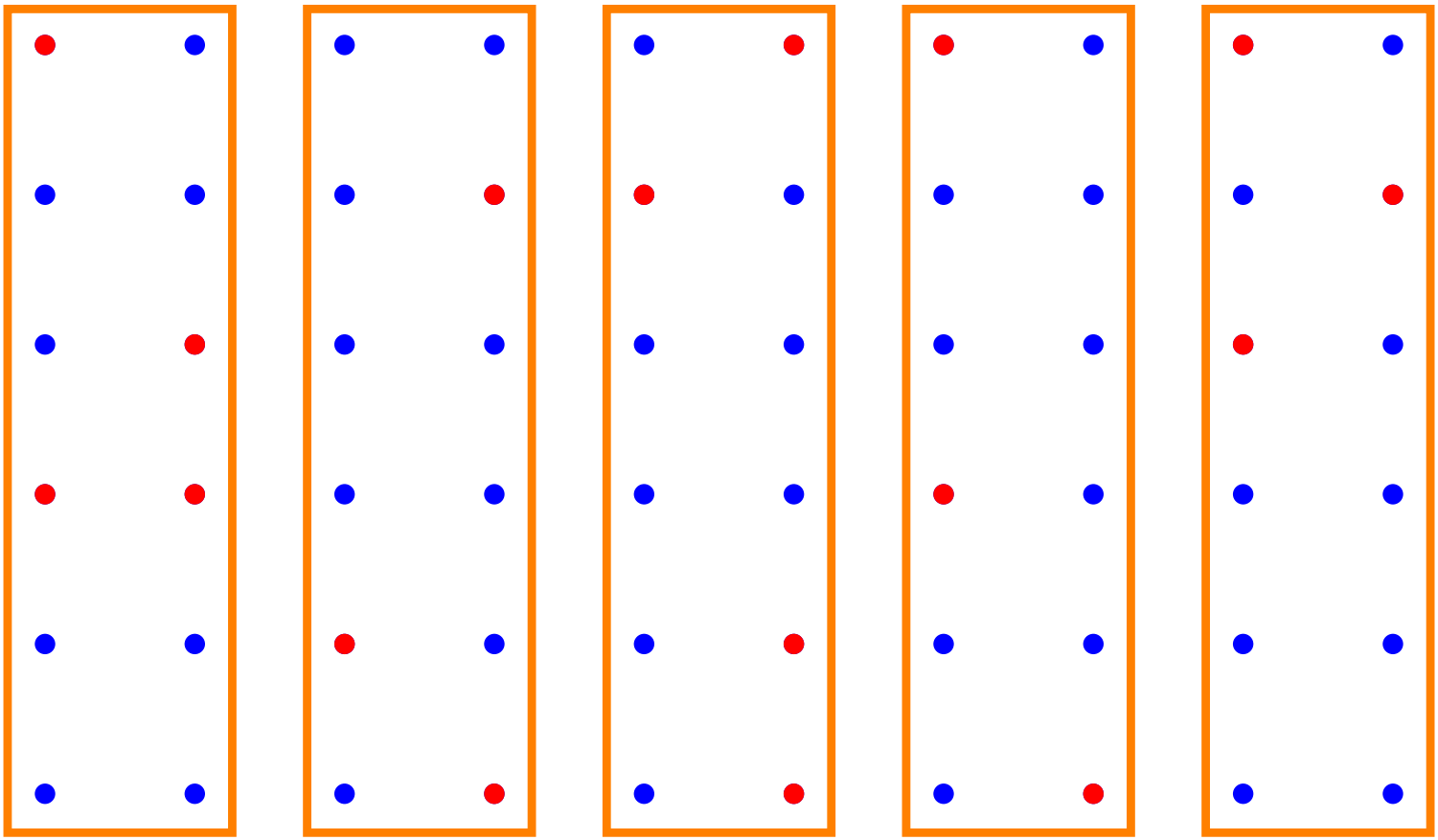
Test parfait : quelle stratégie ?

Prévalence de $p \in]0;1[$, individu $X_i \sim \mathcal{B}(p)$ d'être malade/positif.



Test parfait : quelle stratégie ?

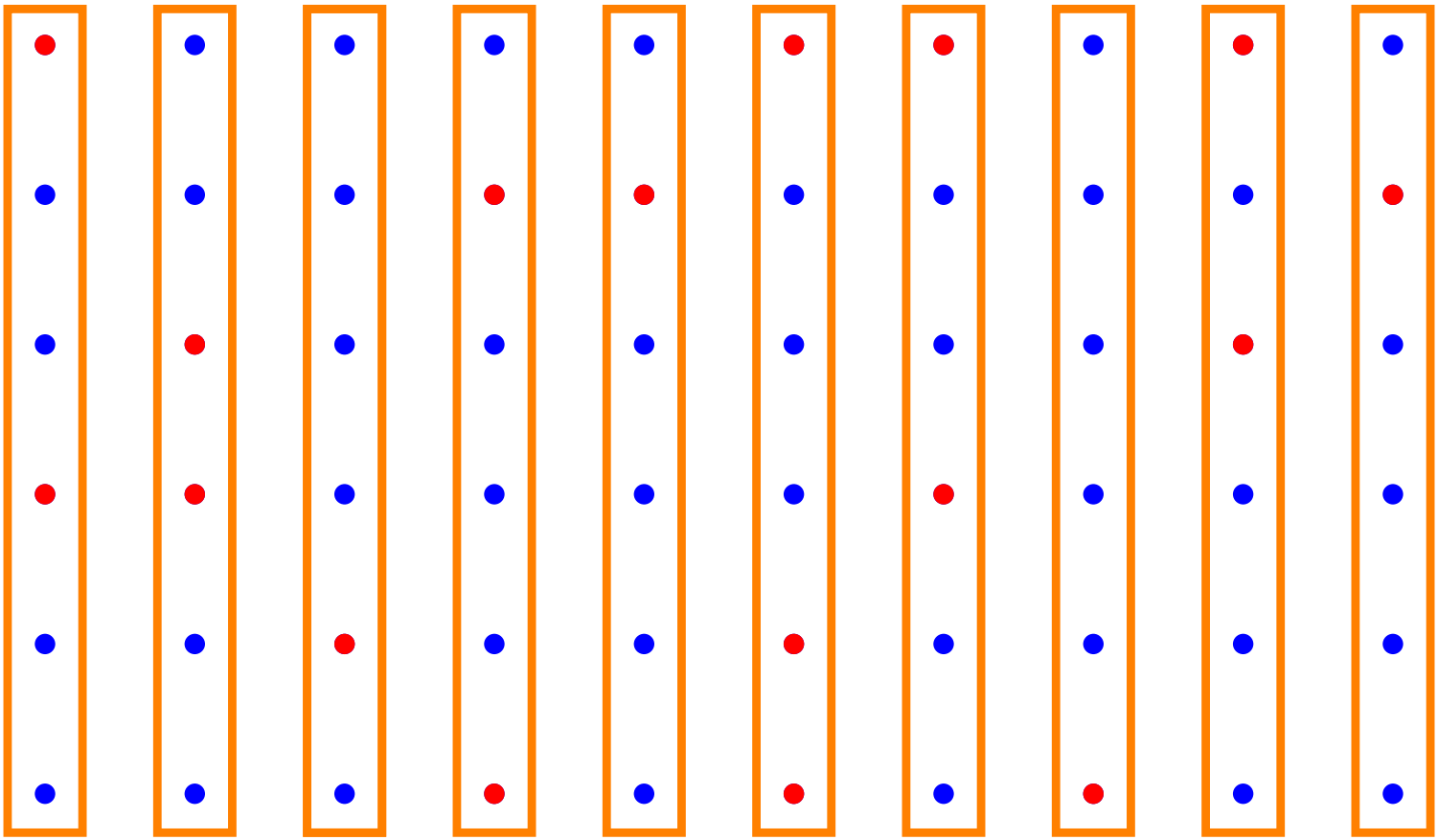
Prévalence de $p \in]0;1[$, individu $X_i \sim \mathcal{B}(p)$ d'être malade/positif.



5 tests groupés + 60 individuels

Test parfait : quelle stratégie ?

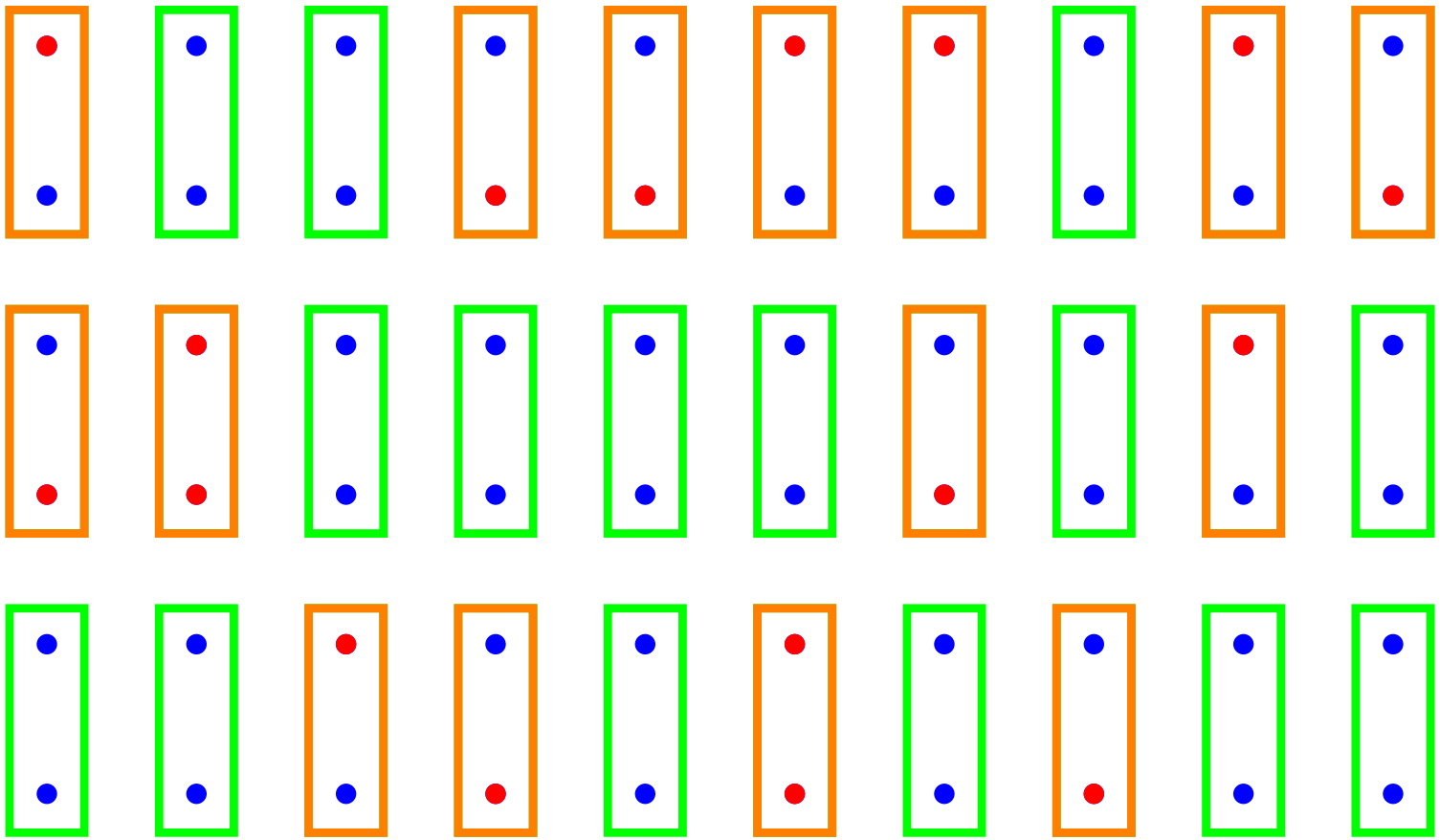
Prévalence de $p \in]0;1[$, individu $X_i \sim \mathcal{B}(p)$ d'être malade/positif.



10 tests groupés + 60 individuels

Test parfait : quelle stratégie ?

Prévalence de $p \in]0;1[$, individu $X_i \sim \mathcal{B}(p)$ d'être malade/positif.



30 tests groupés + 30 individuels

Test parfait : quelle stratégie ?

Prévalence de $p \in]0;1[$, individu $X_i \sim \mathcal{B}(p)$ d'être malade/positif. Groupe de N individus et $X_j^{(N)}$ le résultat du test du groupe j :

$$\mathbb{P}\left(X_j^{(N)} = 1\right) = 1 - (1-p)^N.$$

Test parfait : estimer la prévalence

Avec n groupes de taille N, un estimateur fortement consistant de p est :

$$\widehat{p}_n = 1 - \left(1 - \overline{X}_n^{(N)}\right)^{1/N} \text{ avec } \overline{X}_n^{(N)} = \frac{1}{n}\sum_{j=1}^{n} X_j^{(N)}.$$

1

.

Test parfait : estimer la prévalence

Avec n groupes de taille N, un estimateur fortement consistant de p est :

$$\widehat{p}_n = 1 - \left(1 - \overline{X}_n^{(N)}\right)^{1/N} \text{ avec } \overline{X}_n^{(N)} = \frac{1}{n}\sum_{j=1}^{n} X_j^{(N)}.$$

Un intervalle de confiance de niveau asymptotique $1-\alpha$ est :

$$IC_{1-\alpha}(p) = \left[\widehat{p}_n \pm \frac{q_{1-\alpha/2}\left(1 - \overline{X}_n^{(N)}\right)^{1/N-1}\sqrt{\overline{X}_n^{(N)}\left(1 - \overline{X}_n^{(N)}\right)}}{\sqrt{n}N}\right]$$

avec q_α le quantile d'ordre α d'une gaussienne centrée réduite.

1

.

Test parfait : estimer la prévalence

Avec n groupes de taille N, un estimateur fortement consistant de p est :

$$\widehat{p}_n = 1 - \left(1 - \overline{X}_n^{(N)}\right)^{1/N} \text{ avec } \overline{X}_n^{(N)} = \frac{1}{n}\sum_{j=1}^{n} X_j^{(N)}.$$

Un intervalle de confiance de niveau asymptotique $1-\alpha$ est :

$$IC_{1-\alpha}(p) = \left[\widehat{p}_n \pm \frac{q_{1-\alpha/2}\left(1 - \overline{X}_n^{(N)}\right)^{1/N-1}\sqrt{\overline{X}_n^{(N)}\left(1 - \overline{X}_n^{(N)}\right)}}{\sqrt{n}N}\right]$$

avec q_α le quantile d'ordre α d'une gaussienne centrée réduite.
D'après Thompson [1962], le pré-facteur optimal est obtenu pour :

$$N_{opt} = -\frac{c_\star}{\ln(1-p)}$$

[1] La constante est $c_\star = 2 + W(-2e^{-2}) \approx 1.59$ avec W la fonction de Lambert.

Test parfait : estimer la prévalence

Avec n groupes de taille N, un estimateur fortement consistant de p est :

$$\widehat{p}_n = 1 - \left(1 - \overline{X}_n^{(N)}\right)^{1/N} \text{ avec } \overline{X}_n^{(N)} = \frac{1}{n}\sum_{j=1}^{n} X_j^{(N)}.$$

Un intervalle de confiance de niveau asymptotique $1-\alpha$ est :

$$IC_{1-\alpha}(p) = \left[\widehat{p}_n \pm \frac{q_{1-\alpha/2}\left(1 - \overline{X}_n^{(N)}\right)^{1/N-1}\sqrt{\overline{X}_n^{(N)}\left(1 - \overline{X}_n^{(N)}\right)}}{\sqrt{n}N}\right]$$

avec q_α le quantile d'ordre α d'une gaussienne centrée réduite.
D'après Thompson [1962], le pré-facteur optimal est obtenu pour :

$$N_{opt} = -\frac{c_\star}{\ln(1-p)} \Leftrightarrow (1-p)^{N_{opt}} = e^{-c_\star} \approx 20\%.$$

[1] La constante est $c_\star = 2 + W(-2e^{-2}) \approx 1.59$ avec W la fonction de Lambert.

Test parfait : estimer la prévalence

Avec n groupes de taille N, un estimateur fortement consistant de p est :

$$\widehat{p}_n = 1 - \left(1 - \overline{X}_n^{(N)}\right)^{1/N} \text{ avec } \overline{X}_n^{(N)} = \frac{1}{n}\sum_{j=1}^{n} X_j^{(N)}.$$

Un intervalle de confiance de niveau asymptotique $1-\alpha$ est :

$$IC_{1-\alpha}(p) = \left[\widehat{p}_n \pm \frac{q_{1-\alpha/2}\left(1 - \overline{X}_n^{(N)}\right)^{1/N-1}\sqrt{\overline{X}_n^{(N)}\left(1 - \overline{X}_n^{(N)}\right)}}{\sqrt{n}N}\right]$$

avec q_α le quantile d'ordre α d'une gaussienne centrée réduite.
D'après Thompson [1962], le pré-facteur optimal est obtenu pour :

$$N_{opt} = -\frac{c_\star}{\ln(1-p)} \Leftrightarrow (1-p)^{N_{opt}} = e^{-c_\star} \approx 20\%.$$

[1] La constante est $c_\star = 2 + W(-2e^{-2}) \approx 1.59$ avec W la fonction de Lambert.

▸ Procédure adaptative

Prévalence dans la population

1. Choix d'une première prévalence $p^{(0)}$.

Prévalence dans la population

1. Choix d'une première prévalence $p^{(0)}$.
2. Estimation du nombre N_{opt} optimal sur la base de $p^{(0)}$ qui minimise le nombre total de tests nécessaires pour obtenir l'estimation de la prévalence p avec la précision et la confiance ciblées.

Prévalence dans la population

1. Choix d'une première prévalence $p^{(0)}$.
2. Estimation du nombre N_{opt} optimal sur la base de $p^{(0)}$ qui minimise le nombre total de tests nécessaires pour obtenir l'estimation de la prévalence p avec la précision et la confiance ciblées.
3. Construction d'un nombre de n pools contenant chacun N_{opt} individus sélectionnés au hasard dans la population générale, avec n le nombre de tests disponibles pour la mesure.

Prévalence dans la population

1. Choix d'une première prévalence $p^{(0)}$.
2. Estimation du nombre N_{opt} optimal sur la base de $p^{(0)}$ qui minimise le nombre total de tests nécessaires pour obtenir l'estimation de la prévalence p avec la précision et la confiance ciblées.
3. Construction d'un nombre de n pools contenant chacun N_{opt} individus sélectionnés au hasard dans la population générale, avec n le nombre de tests disponibles pour la mesure.
4. Estimation du nombre moyen de groupes positifs.

Prévalence dans la population

1. Choix d'une première prévalence $p^{(0)}$.
2. Estimation du nombre N_{opt} optimal sur la base de $p^{(0)}$ qui minimise le nombre total de tests nécessaires pour obtenir l'estimation de la prévalence p avec la précision et la confiance ciblées.
3. Construction d'un nombre de n pools contenant chacun N_{opt} individus sélectionnés au hasard dans la population générale, avec n le nombre de tests disponibles pour la mesure.
4. Estimation du nombre moyen de groupes positifs.
5. Amélioration de l'estimation de la prévalence :

$$p^{(1)} = 1 - \left(1 - \overline{X}_n^{(N)}\right)^{1/N_{opt}}.$$

Prévalence dans la population

1. Choix d'une première prévalence $p^{(0)}$.
2. Estimation du nombre N_{opt} optimal sur la base de $p^{(0)}$ qui minimise le nombre total de tests nécessaires pour obtenir l'estimation de la prévalence p avec la précision et la confiance ciblées.
3. Construction d'un nombre de n pools contenant chacun N_{opt} individus sélectionnés au hasard dans la population générale, avec n le nombre de tests disponibles pour la mesure.
4. Estimation du nombre moyen de groupes positifs.
5. Amélioration de l'estimation de la prévalence :

$$p^{(1)} = 1 - \left(1 - \overline{X}_n^{(N)}\right)^{1/N_{opt}}.$$

6. Retour à l'étape 1 jusqu'à stabilisation.

Prévalence dans la population

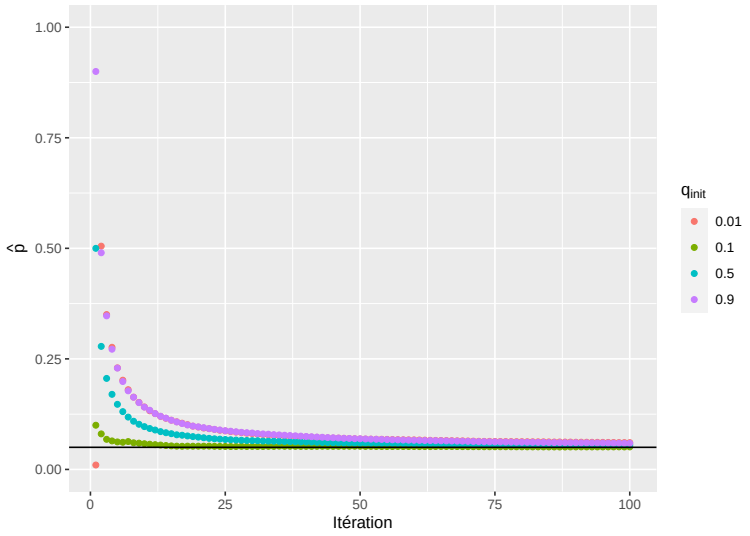
1. Choix d'une première prévalence $p^{(0)}$.
2. Estimation du nombre N_{opt} optimal sur la base de $p^{(0)}$ qui minimise le nombre total de tests nécessaires pour obtenir l'estimation de la prévalence p avec la précision et la confiance ciblées.
3. Construction d'un nombre de n pools contenant chacun N_{opt} individus sélectionnés au hasard dans la population générale, avec n le nombre de tests disponibles pour la mesure.
4. Estimation du nombre moyen de groupes positifs.
5. Amélioration de l'estimation de la prévalence :

$$p^{(1)} = 1 - \left(1 - \overline{X}_n^{(N)}\right)^{1/N_{opt}}.$$

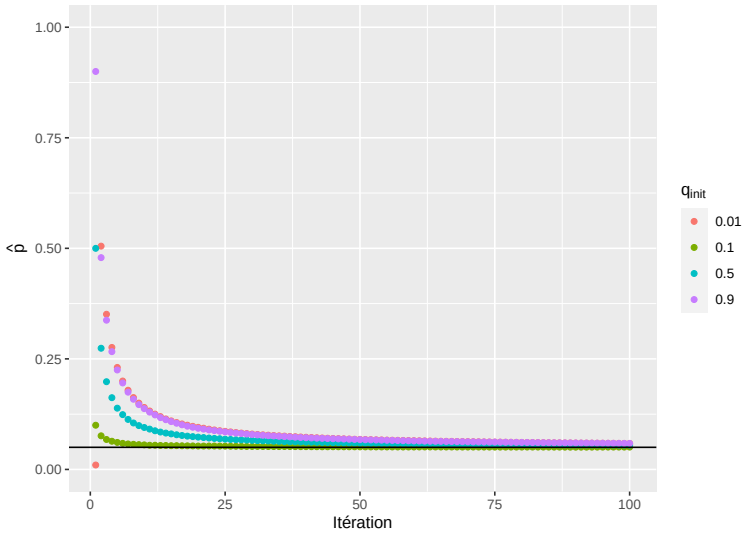
6. Retour à l'étape 1 jusqu'à stabilisation.

Nous pouvons prendre une moyenne des valeurs pour limiter les effets des valeurs extrêmes.

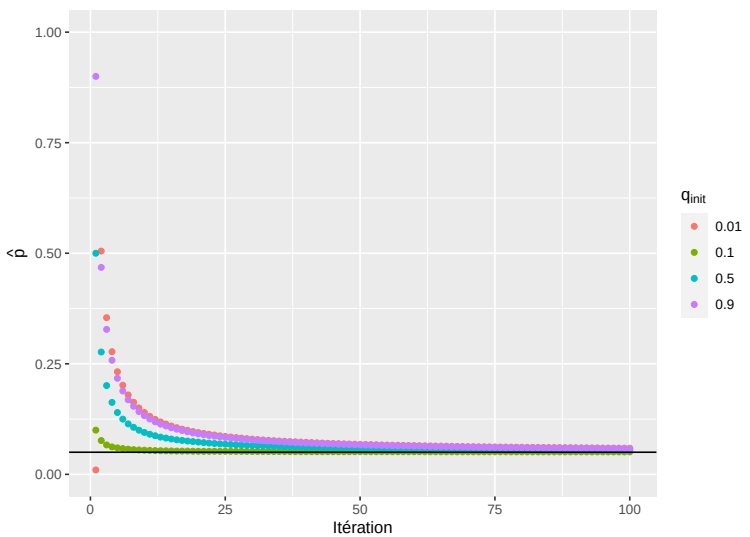
$n = 100$



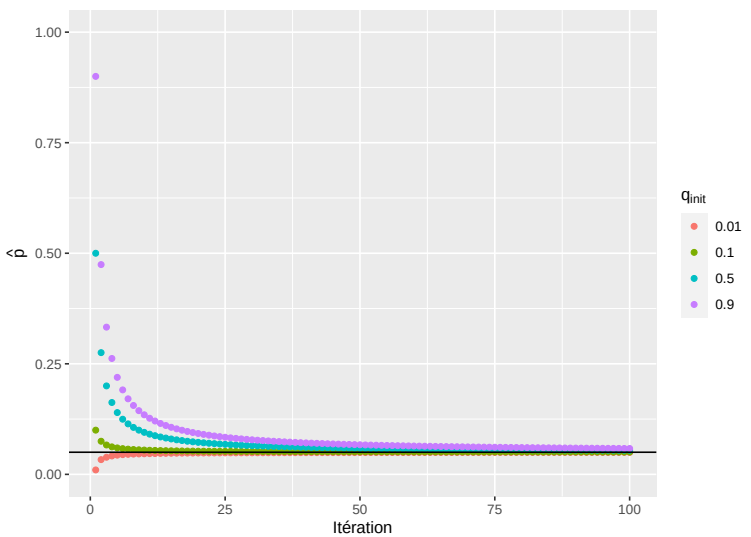
$n = 500$



$n = 1000$

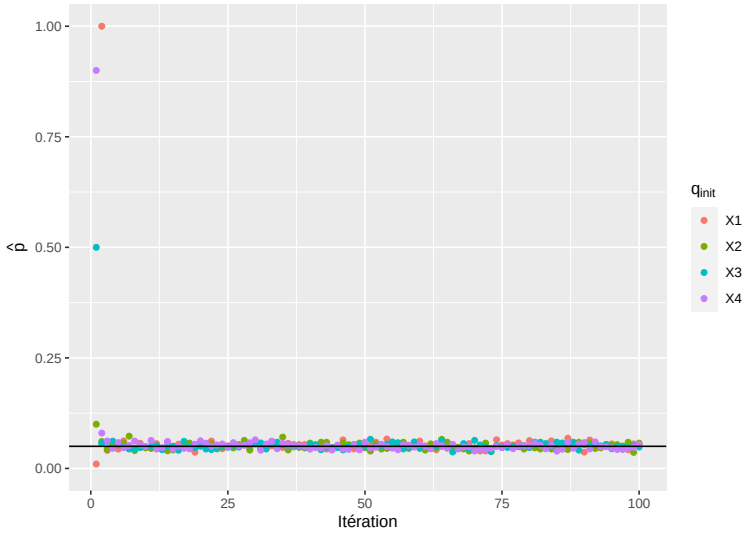


$n = 10000$

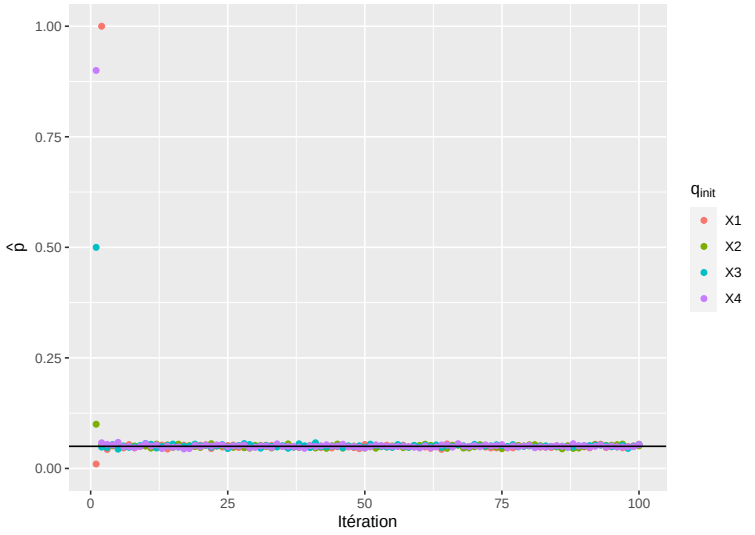


◀ Revenir à l'exposé

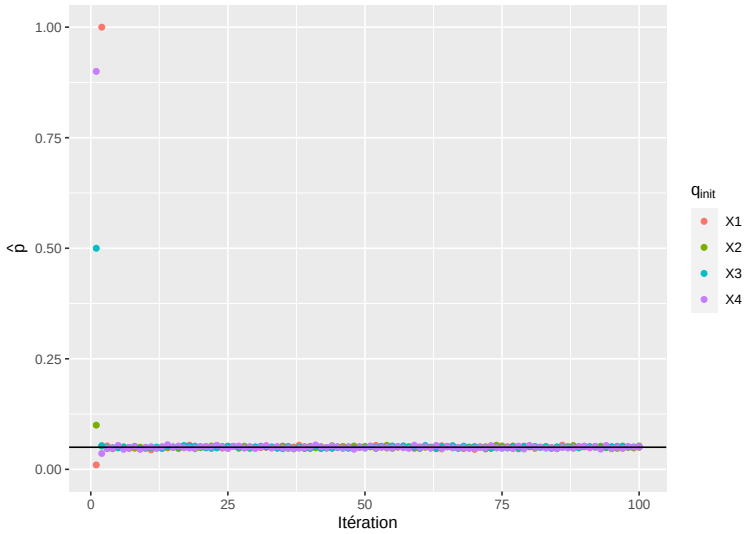
$n = 100$



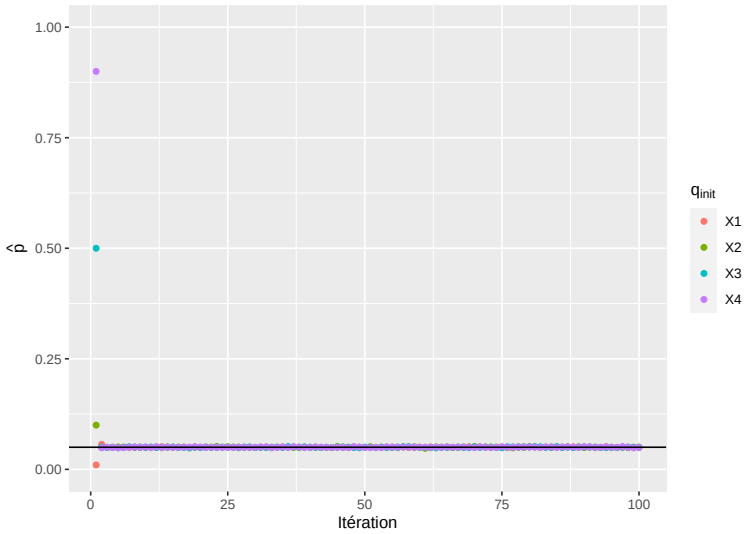
$n = 500$



$n = 1000$

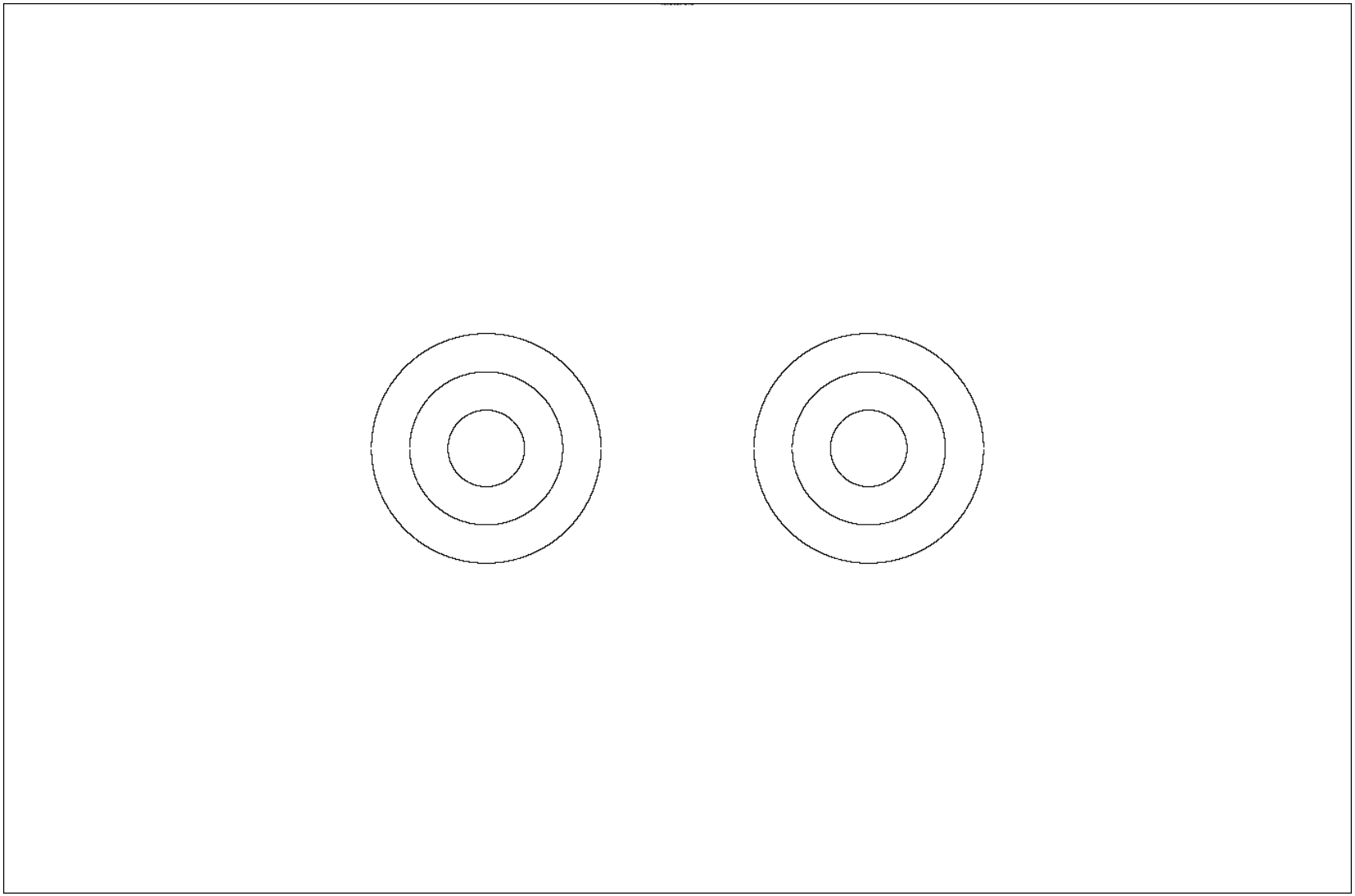


$n = 10000$

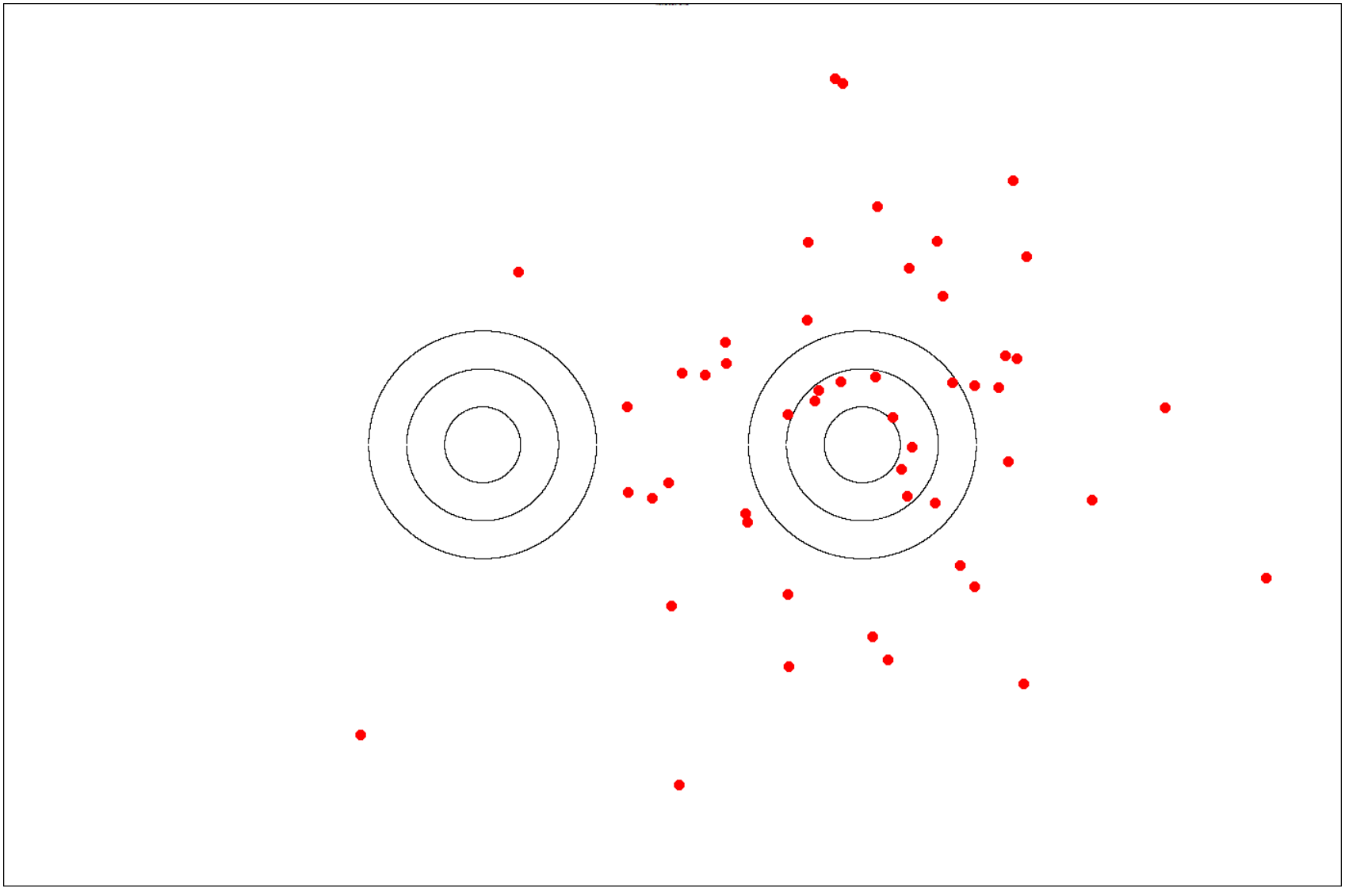


◀ Revenir à l'exposé

Plan



Revenir à l'exposé



Revenir à l'exposé

Loi multinormale

$$Y \sim \mathcal{N}\left(\mu = \begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix}, \sigma^2 I_2\right)$$

La densité est :

$$p(Y; \mu, \sigma^2) = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}\|Y-\mu\|_2^2}$$

Loi multinormale

$$Y \sim \mathcal{N}\left(\mu = \begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix}, \sigma^2 I_2\right)$$

La densité est :

$$p(Y; \mu, \sigma^2) = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}\|Y-\mu\|_2^2}$$

La vraisemblance vaut :

$$\begin{aligned} p(Y_1, \ldots, Y_n; \mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}\|Y_i-\mu\|_2^2} \\ &= \frac{1}{(2\pi\sigma^2)^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\|Y_i-\mu\|_2^2\right) \\ \Rightarrow \log p(Y_1, \ldots, Y_n; \mu, \sigma^2) &= -n\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\|Y_i-\mu\|_2^2 \end{aligned}$$

Estimateurs du maximum de vraisemblance

$$\frac{\partial}{\partial \mu_j} \log p\left(Y_1, \ldots, Y_n; \widehat{\mu}, \widehat{\sigma}^2\right) = 0 \quad \Leftrightarrow \quad \frac{1}{\widehat{\sigma}^2} \sum_{i=1}^{n} (Y_{ij} - \widehat{\mu}_j) = 0$$

$$\Leftrightarrow \quad \widehat{\mu}_j = \frac{1}{n} \sum_{i=1}^{n} Y_{ij}$$

Estimateurs du maximum de vraisemblance

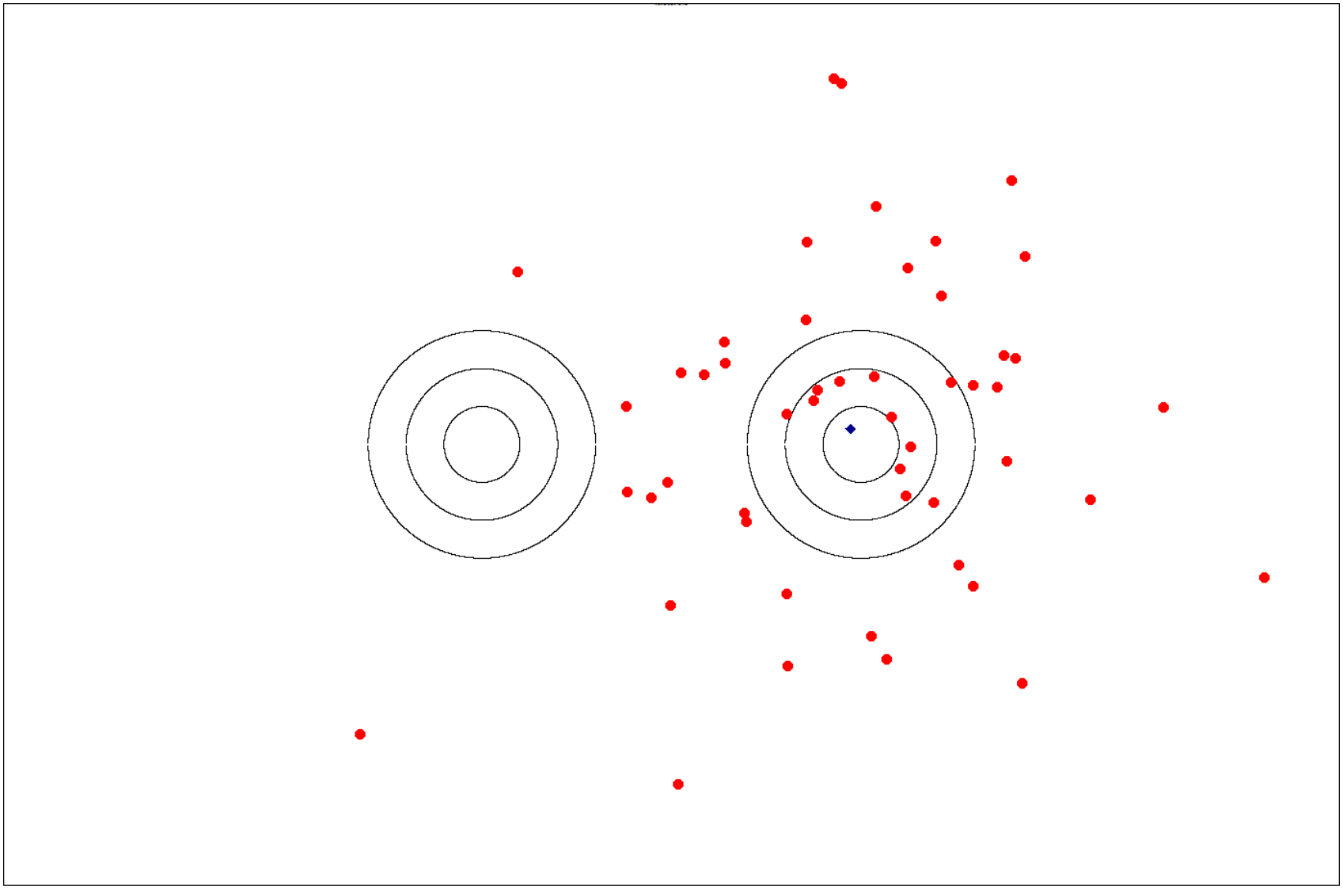
$$\frac{\partial}{\partial \mu_j} \log p\left(Y_1, \ldots, Y_n; \widehat{\mu}, \widehat{\sigma}^2\right) = 0 \quad \Leftrightarrow \quad \frac{1}{\widehat{\sigma}^2} \sum_{i=1}^{n} (Y_{ij} - \widehat{\mu}_j) = 0$$

$$\Leftrightarrow \quad \widehat{\mu}_j = \frac{1}{n} \sum_{i=1}^{n} Y_{ij}$$

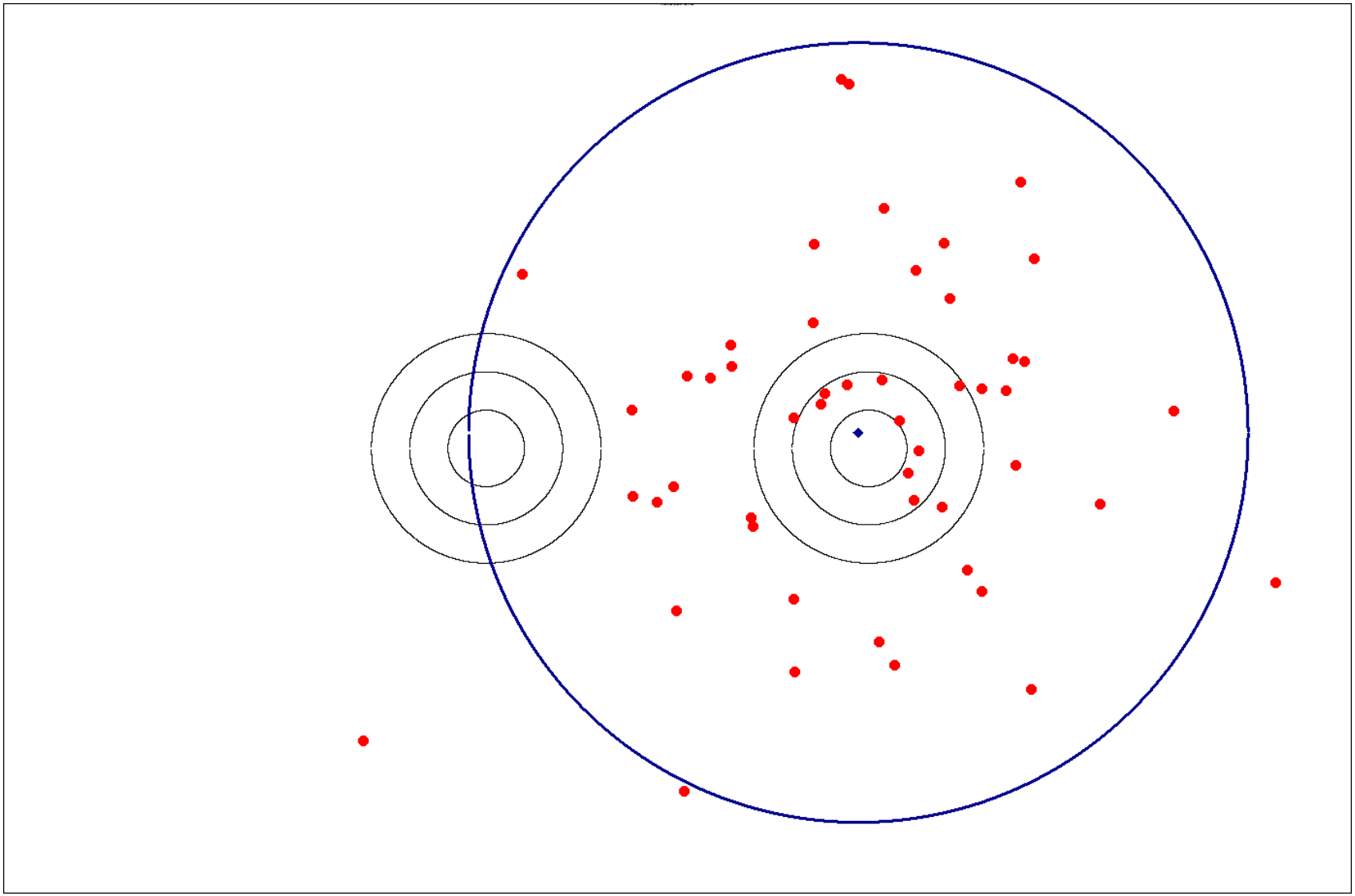
$$\frac{\partial}{\partial \sigma^2} \log p\left(Y_1, \ldots, Y_n; \widehat{\mu}, \widehat{\sigma}^2\right) = 0$$

$$\Leftrightarrow \quad -\frac{n}{\widehat{\sigma}^2} + \frac{1}{2\left(\widehat{\sigma}^2\right)^2} \sum_{i=1}^{n} \|Y_i - \widehat{\mu}\|^2 = 0$$

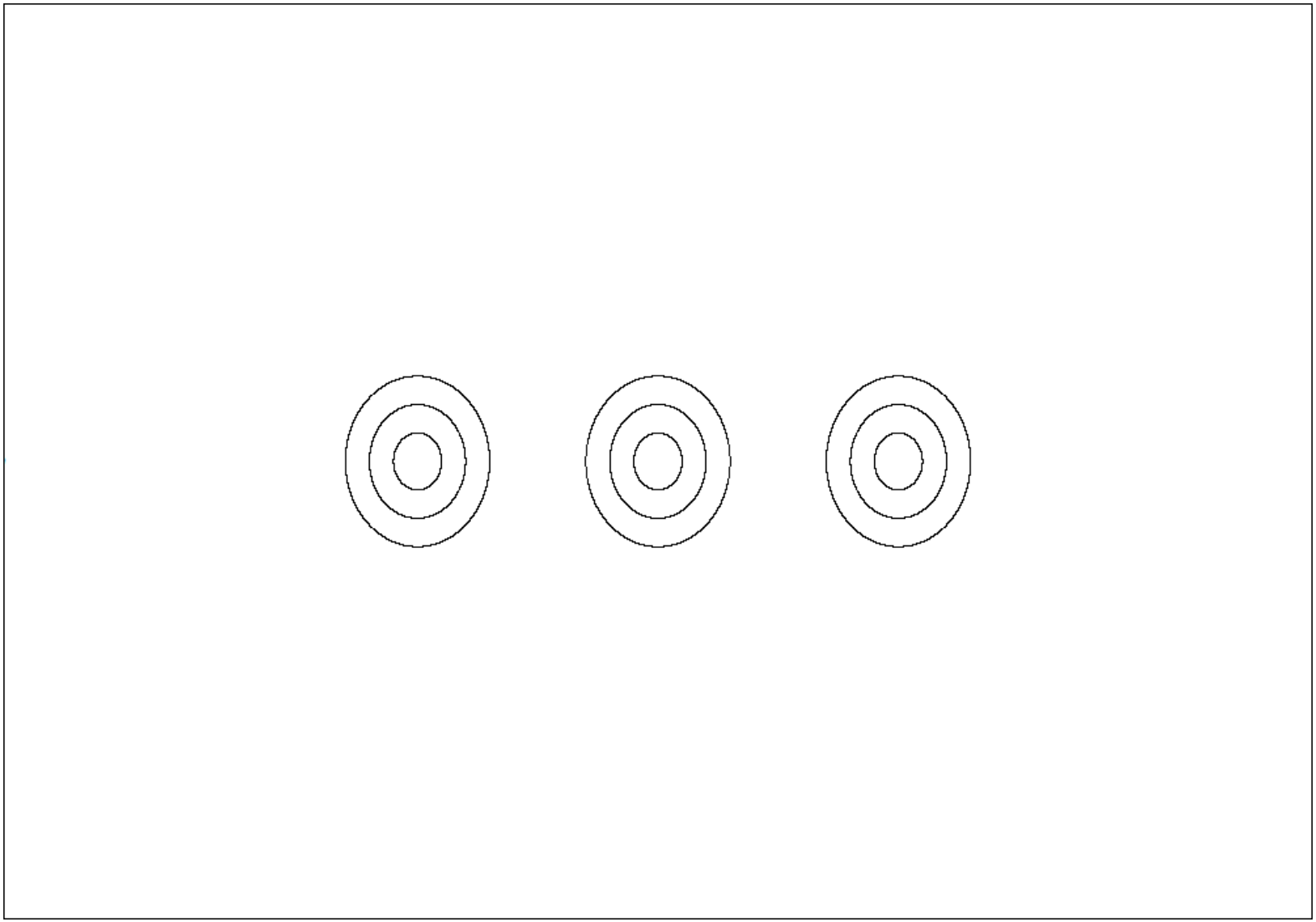
$$\Leftrightarrow \quad \widehat{\sigma}^2 = \frac{1}{2n} \sum_{i=1}^{n} \|Y_i - \widehat{\mu}\|^2$$



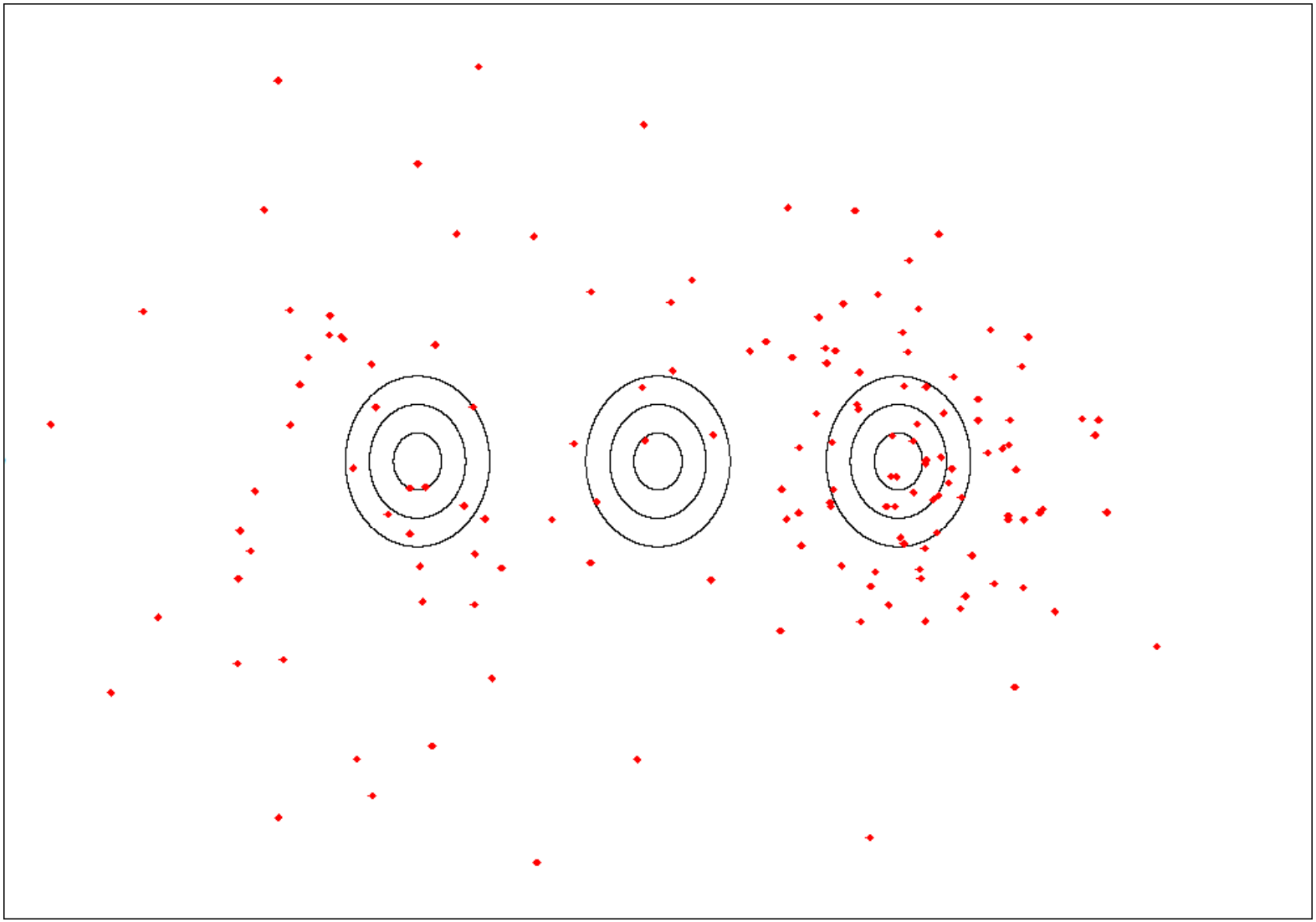
Revenir à l'exposé



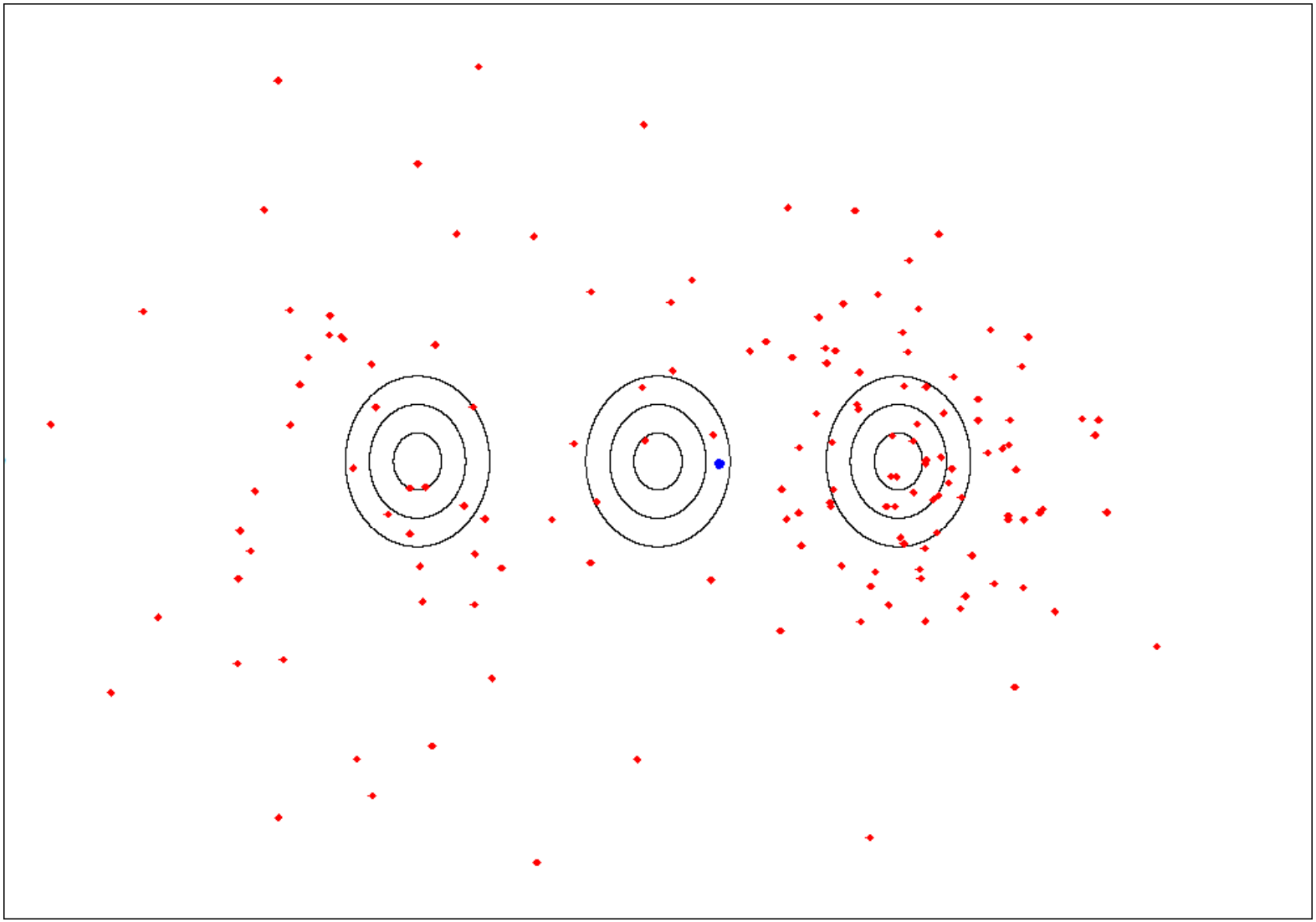
Revenir à l'exposé



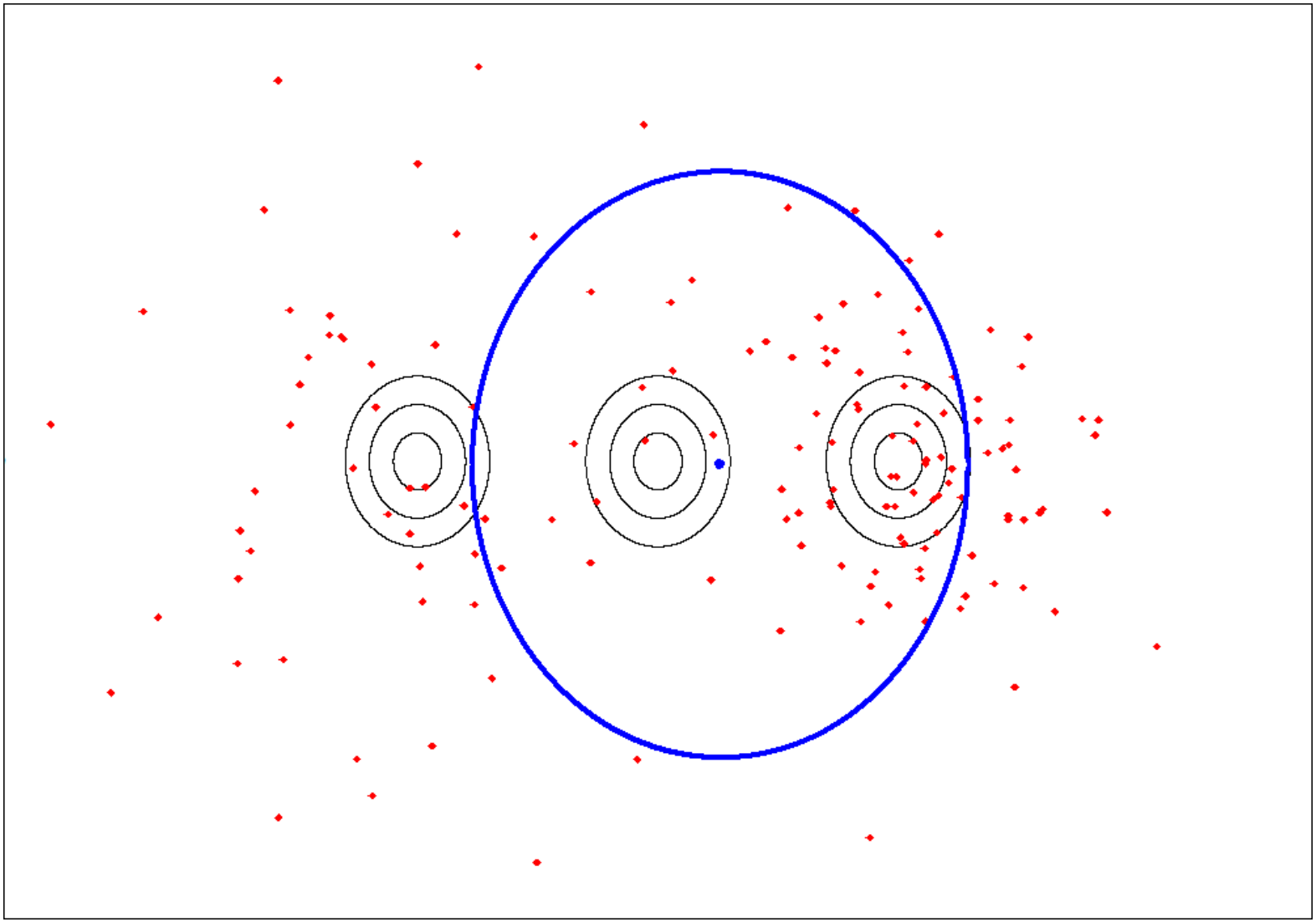
Revenir à l'exposé



Revenir à l'exposé



Revenir à l'exposé



Revenir à l'exposé

Notation

z une matrice $n \times K$ telle que

$$z_{ik} = 1 \Leftrightarrow y_i \text{ appartient à la classe } k$$

et $\mathbb{P}(Z_{ik} = 1) = \pi_k$.
$\varphi(.; \theta_k)$ la densité de la classe k :

$$\mathbb{P}(Y_i | Z_{ik} = 1; \theta) = \varphi(Y_i; \theta_k) = \frac{1}{2\pi\sigma_k^2} e^{-\frac{1}{2\sigma_k^2} ||Y_i - \mu_k||_2^2}$$

◂ Revenir à l'exposé

Modèle de mélange

$$
\begin{aligned}
p(Y_i;\theta) &= \sum_{k=1}^{K} \mathbb{P}(Y_i|Z_{ik}=1;\theta)\mathbb{P}(Z_{ik}=1;\theta) \\
&= \sum_{k=1}^{K} \pi_k \varphi(Y_i;\theta_k)
\end{aligned}
$$

◂ Revenir à l'exposé

Vraisemblance compliquée

$$
\begin{aligned}
L(\theta) &= \log p(Y_1, \ldots, Y_n; \theta) \\
&= \sum_{i=1}^{n} \log p(Y_i; \theta) \\
&= \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k \varphi(Y_i; \theta_k) \right)
\end{aligned}
$$

Revenir à l'exposé

Plan

Idée de la démonstration

Les preuves viennent du fait que la densité appartiennent à la famille des lois exponentielles :

$$f(x) = b(\eta)e^{\langle \eta, T(x) \rangle}$$

avec $< \cdot, \cdot >$ le produit scalaire, $\eta = \left(\frac{\mu}{\sigma^2}, -\frac{1}{\sigma^2}, \ln(q)\right)$ le paramètre naturel et la statistique suffisante :

$$T(x) = \left(x, x^2, \mathbb{1}_{\{x > d_{cens}\}}\right).$$

Pour la constante de normalisation, elle vaut :

$$b(\eta) = \frac{1}{q + (1-q)F_{\mu,\sigma}(d_{cens})} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{mu^2}{2\sigma^2}}.$$

◂ Revenir aux propriétés

Plan

Formule des faux négatifs

Nous avons la densité :

$$f(x) = \sum_{k=1}^{K} \left(\pi_k \frac{f_{\mu_k,\sigma_k}(x)}{q_k + (1-q_k)F_{\mu_k,\sigma_k}(x)} \left[1 + (q_k - 1)\mathbb{1}_{\{x > d_{cens}\}} \right] \right).$$

◂ Revenir aux applications

Formule des faux négatifs

Nous avons la densité :

$$f(x) = \sum_{k=1}^{K} \left(\pi_k \frac{f_{\mu_k,\sigma_k}(x)}{q_k + (1-q_k) F_{\mu_k,\sigma_k}(x)} \left[1 + (q_k - 1) \mathbb{1}_{\{x > d_{cens}\}} \right] \right).$$

Donc, une minoration possible des faux positifs est :

$$\mathbb{P}\left(\text{Faux négatifs}\right) = \sum_{k=1}^{K} \left(\pi_k \left[1 - F_{\mu_k,\sigma_k}(d_{cens}) \right] (1 - q_k) \right)$$

◂ Revenir aux applications

Formule des faux négatifs

Nous avons la densité :

$$f(x) = \sum_{k=1}^{K} \left(\pi_k \frac{f_{\mu_k,\sigma_k}(x)}{q_k + (1-q_k) F_{\mu_k,\sigma_k}(x)} \left[1 + (q_k - 1) \mathbb{1}_{\{x > d_{cens}\}} \right] \right).$$

Donc, une minoration possible des faux positifs est :

$$\mathbb{P}\left(\text{Faux négatifs}\right) = \sum_{k=1}^{K} \left(\pi_k \left[1 - F_{\mu_k,\sigma_k}(d_{cens}) \right] (1 - q_k) \right)$$

Il faut faire attention à ce que représentent les π_k...

◂ Revenir aux applications

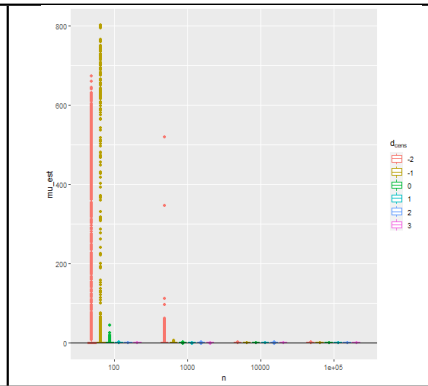
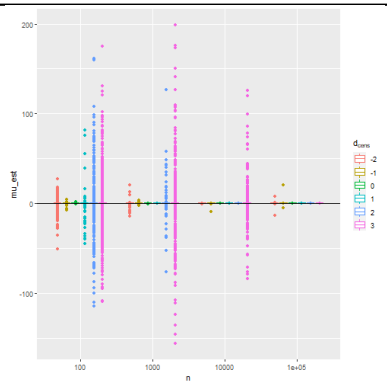
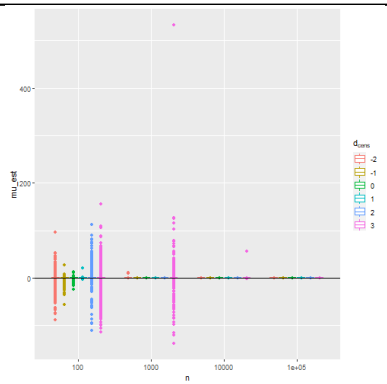
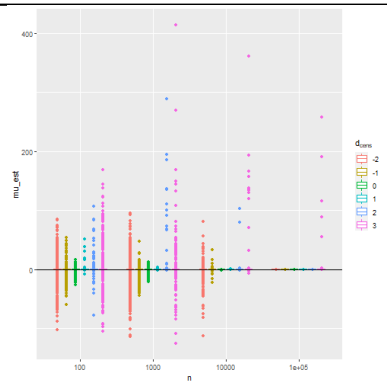
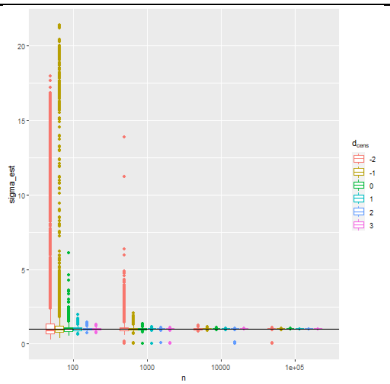
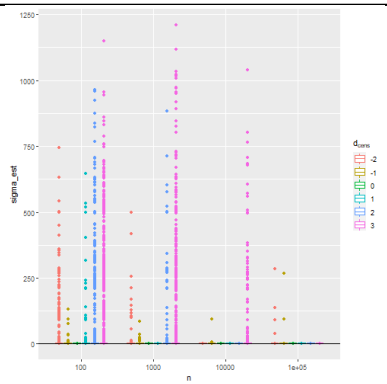
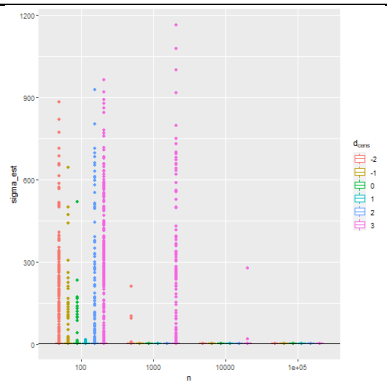
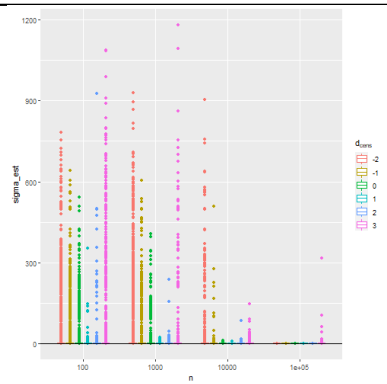
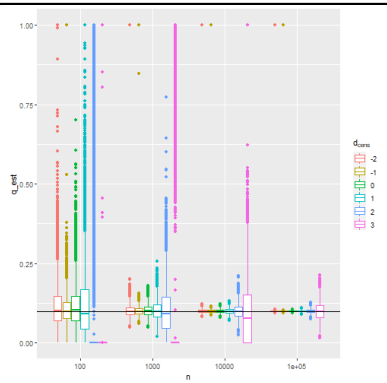
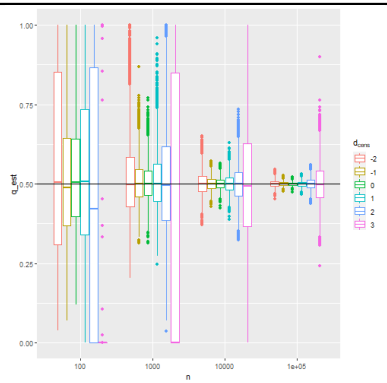
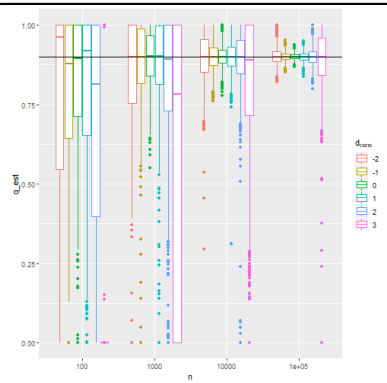
Plan

Estimateur du maximum de vraisemblance

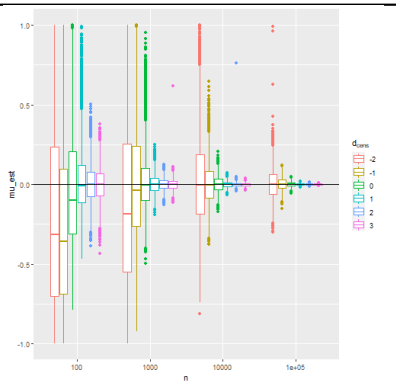
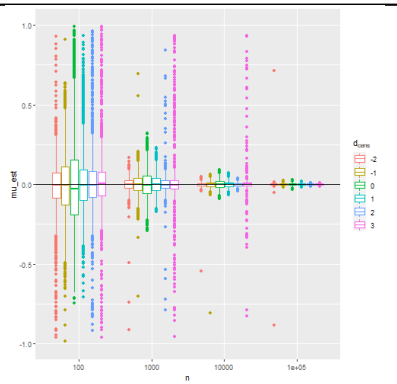
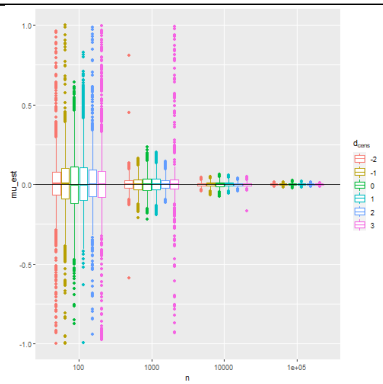
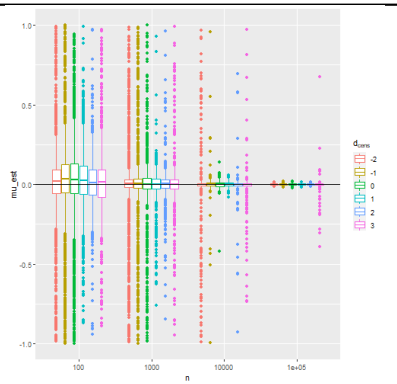
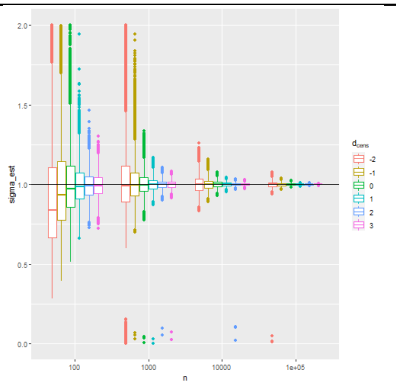
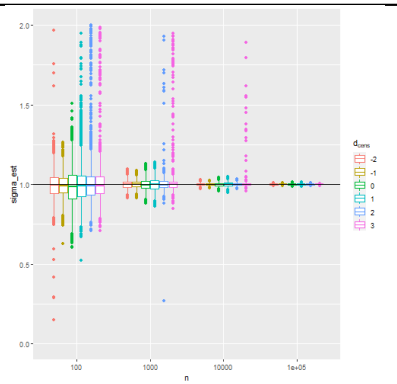
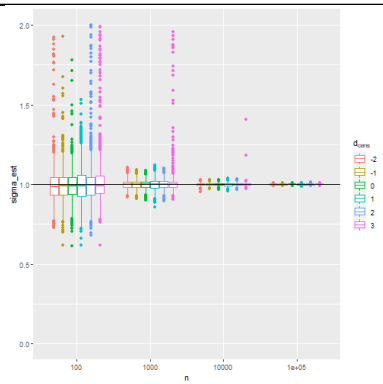
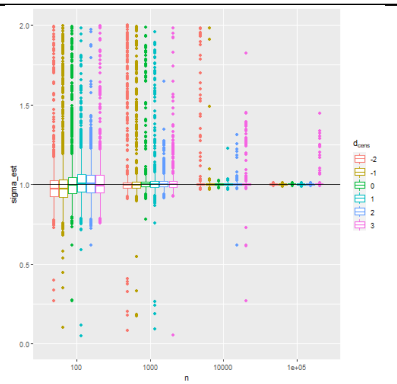
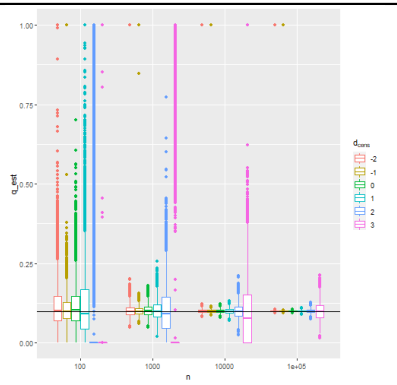
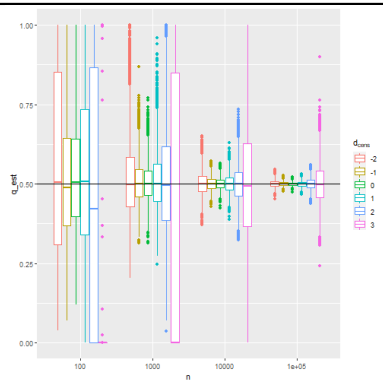
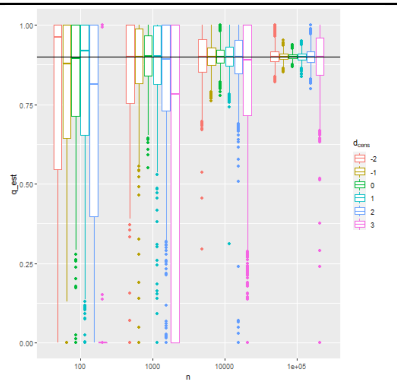
Simulation de 10 000 n-échantillons $\mathcal{CN}_{d_{cens}}(0, 1, q)$ avec :

- $n \in \{10^2, 10^3, 10^4, 10^5\}$,
- $d_{cens} \in \{-2, -1, 0, 1, 2, 3\}$,
- $q \in \{0, 0.1, 0.5, 0.9\}$.

◂ Revenir à la conclusion

	Complète $q = 0$	Partielle $q = 0.1$	Partielle $q = 0.5$	Partielle $q = 0.9$
$\widehat{\mu}$				
$\widehat{\sigma}$				
$\widehat{q}$				

Revenir à la conclusion

	Complète $q = 0$	Partielle $q = 0.1$	Partielle $q = 0.5$	Partielle $q = 0.9$
$\widehat{\mu}$				
$\widehat{\sigma}$				
$\widehat{q}$				

Revenir à la conclusion