Echantillonage Décisionnel - README

abelmasson

February 2025

1 Introduction

L'échantillonnage est une méthode consistant à sélectionner un sous-ensemble d'observations (= un échantillon) représentatif d'une population plus large pour répondre à une question. Selon cette question, différents critères peuvent guider la construction de l'échantillon. Dans la plupart des cas, un échantillon de grande taille est préférable, parce qu'il améliore la représentativité de la population et la précision des réponses. Toutefois, dans certaines situations, notamment lorsque l'acquisition de chaque observation est coûteuse, on cherchera plutôt à limiter la taille de l'échantillon. L'échantillonnage décisionnel (ou séquentiel) est particulièrement adapté à ces situations en ce qu'il repose sur une collecte progressive des observations, une par une, afin d'obtenir un échantillon de taille strictement suffisante pour fournir une réponse fiable.

Exemple d'application à l'infestation d'une serre par un insecte

On cherche à surveiller le risque d'infestation d'une serre de fraises en constituant un échantillon de plants de fraisier. Chaque plant est consciencieusement examiné par un technicien pour déterminer la présence ou l'absence de l'insecte. Dans ce contexte, un faible nombre de detection peut être le signal d'une infestation, alors que l'insecte recherché est généralement rare et souvent absent. Ainsi, le nombre de plants à examiner avant de pouvoir statuer sur l'infestation de la serre peut rapidement devenir ingérable. Ce script propose une méthode d'échantillonnage décisionnel visant à optimisiser, c'est-à-dire à réduire au maximum le nombre de plants à examiner.

1.1 Formulation

1.1.1 Formulation du modèle

Pour chaque lot de plants (i.e. chaque serre), on cherche à determiner si la fréquence de l'insecte dépasse un certain seuil au delà duquel on peut considérer que la serre est infestée. On note p la probabilité de retrouver un insecte sur un plant, et $p*$ le seuil critique à ne pas dépasser (par exemple 1 % des plants soit $p* = 0.01$). On suppose que p suit une loi $Beta$ et que X le nombre de

plants infestés suit une loi Binomiale de paramètres N le nombre total de plants examinés et p. Ainsi :

$$X \sim Binomial(N, p)$$

avec

$$p \sim Beta(\alpha, \beta)$$

1.1.2 Formulation du test

On pose **(H)** l'hypothèse : **"la probabilité p de trouver un insecte sur un plant dépasse le seuil critique $p*$"**

On cherche à tester **(H)** en fonction des résultats des examens de plants, c'est-à-dire en fonction des observations de X. Pour ce faire on se munit d'un seuil de confiance e permettant d'accepter l'hypothèse **(H)** (lorsque $P(Hvraie) > 1-e$), de la rejeter (lorsque $P(Hvraie) < e$), ou de rester indécis si les résultats des examens ne permettent pas de prendre une décision fiable (lorsque $1 - e > P(Hvraie) > e$).

Rappel sur la loi Beta

La loi $Beta$ est une distribution continue définie par deux paramètres α et β. Elle est classiquement utilisée pour représenter des proportions ou des probabilités. En particulier lorsque l'on s'interesse à des données distribuées selon une loi binomiale ($X \sim Binomial(N, p)$), la loi $Beta$ est adapté à la représentation de la probabilité p pour plusieurs raisons :

- La loi $Beta$ est définie sur l'intervalle $[0, 1]$.
- La loi $Beta(1, 1)$ est identique à la loi uniforme sur $[0, 1]$ et constitue une *prior* adaptée lorsqu'aucune autre information sur p n'est disponible.
- Lorsque des observations de X sont disponibles, il existe une expression analytique pour en informer la distribution de p de façon simple et intuitive : La *posterior* de p sera la loi $Beta(1 + \alpha, 1 + \beta)$ avec α et β les nombres de succès et d'echecs observés. On dit que la loi $Beta$ est la conjuguée de la loi Binomiale.

1.2 Protocole d'échantillonage décisionnel

Comme expliqué en introduction, le protocole d'échantillonage décisionnel repose sur l'intègration des observations une par une. Plus précisement il reestime la distribution de probabilité de p et reitere le test sur (H) à chaque nouvel examen de plant. Le détail du protocole est le suivant :

- On part sans information a priori sur la distributiuon de p, soit $p \sim Beta(1, 1)$

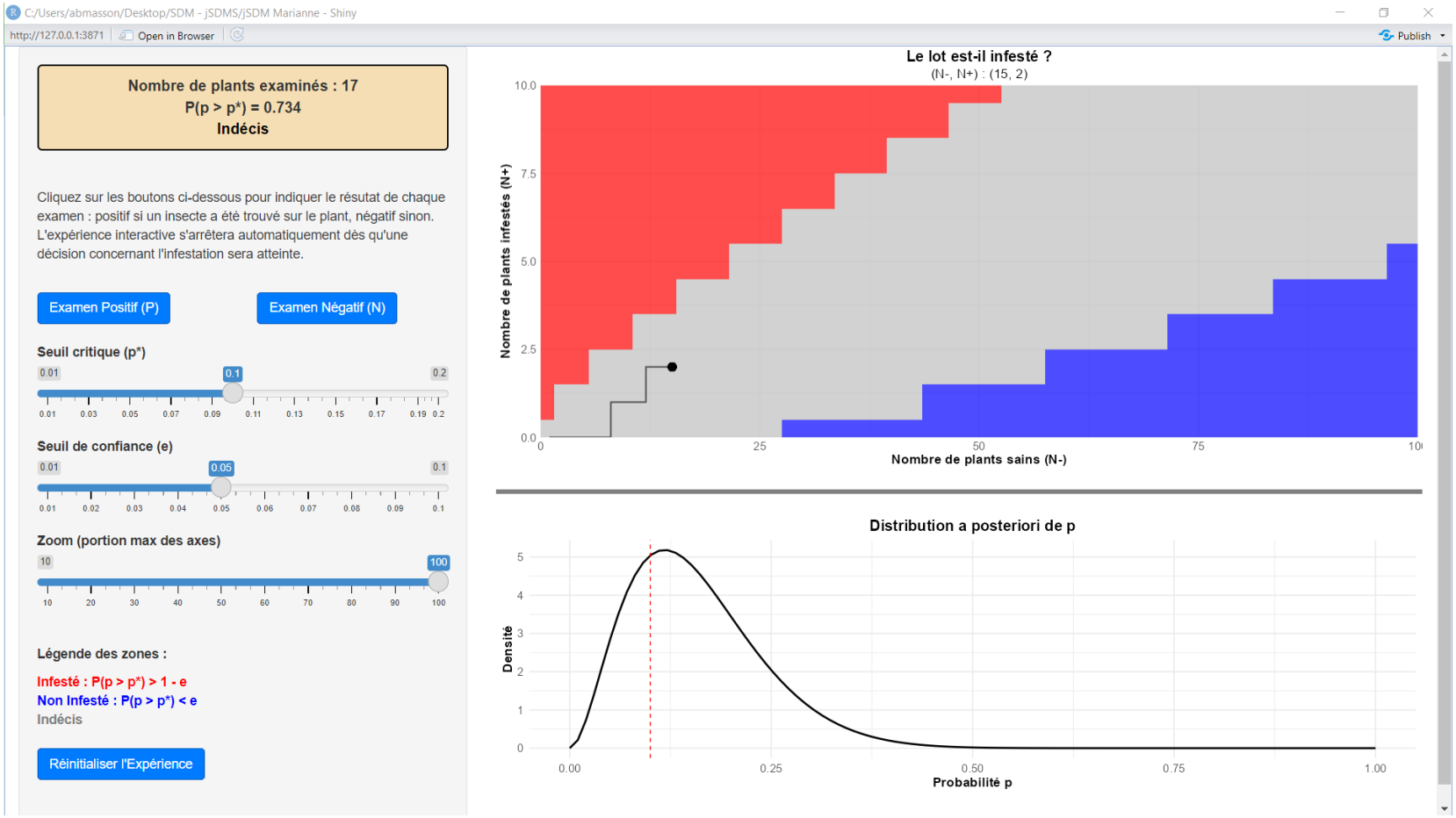
- A chaque nouvel examen de plant :
  - On actualise la distribution de p, en ajoutant 1 à α si le test est positif, 1 à β s'il est négatif (voir rappel sur la loi Beta).
  - On calcule la probabilité P(H vraie) que p dépasse le seuil critique $p*$.
    - Si P(H vraie) est supérieure à $1 - e$ alors on conclut que **le lot est infesté** et on s'arrête.
    - Si P(H vraie) est inférieure à e on conclut que **le lot n'est pas infesté** et on s'arrête.
    - Si P(H vraie) est comprise entre e et $1 - e$ enfin, un **examen supplémentaire est nécessaire**, on recommence la boucle avec un nouvel examen.
- Si l'on atteint le nombre de plants dans le lot avant d'avoir pu conclure que le lot est ou n'est pas infesté, on conclut d'une **indécision**.

2 Expérience Interactive

Le code .Rmd présenté dans ce *reposiroty* contient un RShiny pour implémenter et viusualiser le protocole d'échantillonage décisionnel. Ci-après une capture d'écran de l'interface d'expérience interactive (Figure 1). Elle en donne une représentation graphique comprenant :

- Un graphique avec en ordonnées le nombre de plant examinés infestés (N_+) et en abscisse le nombre de plants examinés saints (N_-). Sur ce graphique on représente en rouge la zone où $P(Hvraie|X = \frac{N_+}{N_+ + N_-})$ $1 - e$ et en bleu $P(Hvraie|X = \frac{N_+}{N_+ + N_-}) > 1 - e$
- Un graphique donnant la distribution de probabiltié p qui correspond à une distribution $Beta(1 + N_+, 1 + N_-)$.

L'experience interactive prend automatiquement fin dès qu'une décicision a été atteinte.


C:/Users/abmasson/Desktop/SDM - jSDMS/jSDM Marianne - Shiny
http://127.0.0.1:3871
Open in Browser
Publish
Nombre de plants examinés : 17
P(p > p*) = 0.734
Indécis
Cliquez sur les boutons ci-dessous pour indiquer le résutat de chaque examen : positif si un insecte a été trouvé sur le plant, négatif sinon. L'expérience interactive s'arrêtera automatiquement dès qu'une décision concernant l'infestation sera atteinte.
Examen Positif (P)
Examen Négatif (N)
Seuil critique (p*)
0.01
0.1
0.2
0.01 0.03 0.05 0.07 0.09 0.11 0.13 0.15 0.17 0.19 0.2
Seuil de confiance (e)
0.01
0.05
0.1
0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.1
Zoom (portion max des axes)
10
100
10 20 30 40 50 60 70 80 90 100
Légende des zones :
Infesté : P(p > p*) > 1 - e
Non Infesté : P(p > p*) < e
Indécis
Réinitialiser l'Expérience
Le lot est-il infesté ?
(N-, N+) : (15, 2)
Nombre de plants infestés (N+)
10.0
7.5
5.0
2.5
0.0
0
25
50
75
Nombre de plants sains (N-)
Distribution a posteriori de p
Densité
5
4
3
2
1
0
0.00
0.25
0.50
0.75
1.00
Probabilité p