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 afin d'en tirer des conclusions. Selon l'objectif de l'étude, différents critères peuvent guider la construction de l'échantillon. Dans la plupart des cas, un échantillon de grande taille est préférable, car il permet d'améliorer la précision des analyses et d'assurer une meilleure représentativité de la population étudiée. Toutefois dans certaines situations, on cherche plutôt à obtenir l'échantillon le plus petit possible, notamment lorsque l'échantillonnage est destructif ou que le traitement des observations est très coûteux. L'échantillonnage décisionnel (ou séquentiel) en particulier, consiste à collecter les observations une par une, jusqu'à ce que l'échantillon soit de taille suffisante pour prendre décision fiable (à un seuil de certitude donné).

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

On cherche à surveiller le risque d'infestation d'une serre de fraise en constituant un échantillon de plants de fraisiers, et en vérifiant sur chaque plant si l'insecte recherché est présent ou non. Dans le contexte des infestations, même une faible fréquence d'occurence des insectes peut indiquer un risque sérieux. Ainsi le nombre de plants à examiner peut rapidement atteindre une taille très importante, et nécessiter un examen long et fastidieux. Dans ce script, nous développons une méthode d'échantillonage décisionnel permettant de minimiser le nombre de plants de fraisiers à examiner avant de pouvoir statuer sur l'infestation de la serre.

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 souvent utilisée pour représenter des proportions ou des probabilités. Typiquement lorsque l'on s'interesse à des données distribuées selon une loi binomiale ($X \sim Binomial(N, p)$), la loi $Beta$ est un choix classique pour représenter la probabilité p pour plusieurs raisons :

- Elle est définie sur l'intervalle $[0, 1]$.
- La loi beta $Beta(1, 1)$ est la loi uniforme sur $[0, 1]$. Dans un contexte bayésien, elle constitue une prior adaptée pour p lorsque qu'aucune autre information n'est disponible.
- Dans un contexte bayésien, l'estimation de la posterior d'une variable p suivant une loi Beta de paramètres α et β sachant le résultat d'une épreuve de Bernoulli de probabilité p est simple : p suit toujours une loi $Beta$, de paramètres $\alpha + 1$ et β si le résultat de l'épreuve est positif, α et $\beta + 1$ s'il est négatif. Ainsi lorsque l'on dispose d'observations d'une variable $X \sim Binomial(N, p)$, les paramètres α et β de la posterior de p représentent respectivement le nombre de succès et d'échecs observés, ce qui rend la calibration de p simple (analytique) et intuitive. On dit que la loi beta est la loi 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 que p dépasse le seuil critique $p*$. Si cette probabilité est supérieure à $1 - e$ alors on conclut que **le lot est infesté**, si elle est inférieure à e on conclut que **le lot n'est pas infesté**. Si elle est comprise entre e et $1 - e$ enfin, on ne peut pas conclure et un **examen supplémentaire** est **nécessaire**.

Si l'on atteint le nombre de plants dans le lot, on conclut sur une **indécision**.

2 Expérience Interactive

Le code .Rmd présenté dans ce *reposiroty* contient un RShiny pour implémenter et viusualiser le protocole d'échantillonage décisionnel. Il en donne une représentation graphique comprenant en particulier :

- 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.