Analyse de données censurés

Par

Amine BOUSFIHA

Samia MESSAOUDI

Encadré par : Mme Aurélie CATEL

TABLE DES MATIERES

INTRODUCTION

Dans les cours de statistiques de première et deuxième année de l'ENSIMAG, nous considérons des données non censurées, c'est-à-dire complétement et parfaitement observées. Le problème est qu'en pratique, il est fréquent que les données recueillies soient incomplètes. Ces données sont dites censurées. Par exemple, lors de l'étude de l'influence d'un traitement sur le cancer, il se peut qu'un patient quitte l'étude avant son décès.

L'objectif du projet est d'adapter les méthodes de base de la statistique (estimation, intervalles de confiance, tests) et les implémenter en R, afin de pouvoir automatiser l'analyse d'échantillons de données censurées.

OBJECTIFS

OBJECTIFS TECHNIQUES

L'objectif final sera de rendre un script à exécuter sur un terminal et qui prend en entrée un fichier texte contenant des données censurées ou complètes. Le fichier texte donné en entrée doit être sous une forme prédéfinie. En effet, ce fichier sera formé de deux colonnes ; la première regroupe l'ensemble des observations à analyser, la deuxième colonne nous informe de la nature de l'observation associée : 0 pour une observation complète, 1 si l'observation a été censurée.

Ce script affiche ensuite l'analyse des données en entrée : Une première ligne nous informe si les données en entrée sont censurées, ensuite vient une analyse des différents paramètres associés à la loi exponentielle et à la loi de Weibull. Enfin, une conclusion de cette analyse dépendamment des résultats obtenus sera fournie comme résultat final.

OBJECTIFS DE DELAIS

Au vu de l'analyse des objectifs du projet et des différentes contraintes qui y sont liées, nous avons déterminé les dates clés intermédiaires citées ci-dessous. L'objectif sera de respecter ces dates et de de réorganiser en cas de retard sur une date particulière, ce qui nous permettra de rendre le produit final dans les délais impartis.

MACRO-PLANNING

ETAPES PRINCIPALES DU PROJET

Notre compréhension du sujet nous a permis de le décomposer en plusieurs étapes, que nous avons classées dans l'ordre suivant :

- Analyse de données complètes (non censurées)
- Analyse de données censurées :
  - Censure de Type I
  - Censure de type II
  - Multi-censure

L'analyse de données (complètes ou censurées) se fait en plusieurs étapes :

- Estimation des paramètres : Lambda λ / Eta η & Beta β
- Evaluation de la qualité des estimations : Biais, variance
- Choix du modèle : Exponentielle(λ)/ Weibull(η, β)
- Graphe de probabilité/ Tests d'adéquation, Qualité

DATES CLES

Afin de rendre le produit au client dans les délais imposés, il a été nécessaire d'établir un planning prévisionnel, regroupant les étapes principales du projet citées précédemment, et qui représentes pour nous des rendus intermédiaires :

- Lundi 25 mai: Programme fonctionnel pour l'analyse de données complètes
- Vendredi 29 mai: Programme fonctionnel pour l'analyse de donnée censurées de type I
- Mardi 2 Juin: Programme fonctionnel pour l'analyse de donnée censurées de type II
- Lundi 8 Juin: Programme fonctionnel pour l'analyse de données de type III
- Mercredi 10 Juin: Rapport complet
- Jeudi 11 Juin: Remise du rapport
- Vendredi 12 Juin: Soutenanance

RESSOURCES ET COMMUNICATION

RESSOURCES

Nous disposons pour ce projet de deux ordinateurs portables, ainsi qu'un accès aux ordinateurs de l'Ensimag du Lundi au Vendredi. De plus, nous avons créé un dépôt Git sur le site gitHub.com, ce qui nous permettra de travailler à distance et de synchroniser notre travail automatiquement.

COMMUNICATION

Nous travaillerons principalement dans les locaux de l'Ensimag, et chez un membre de l'équipe pendant les week-ends en cas de retard sur le planning prévisionnel. Nous utilisons Facebook et Whatsapp pour communiquer et se fixer des rendez-vous, et le dépôt Git pour mettre en commun notre travail.

RISQUES ET DEPENDANCES

RISQUES MAJEURS ET ACTIONS ASSOCEES

Risque	Action Associée
Incompréhension de nouveaux concepts	Documentation
Retard sur un rendu intermédiaire	Week-end de travail supplémentaire pour rattraper un éventuel retard
Manque de motivation	Une demi-journée de repos

DEPENDANCES

Nous avons besoin au cours de ce projet de bien comprendre et cerner le sujet traité. Pour cela, une documentation est nécessaire. De ce fait, malgré les innombrables ressources disponibles sur internet, nous sommes dépendants des documents fournis par notre professeur encadrant.

Les rendus intermédiaires doivent être effectués dans l'ordre chronologique. En effet, toute modification de ce calendrier entrainerait une perte de temps, qui aura pour conséquence directe un retard sur le rendu final.