**Projet de Spécialité**

**Analyse statistique de données censurées**

**Introduction**

Dans les cours de statistique 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 recueillis soit incomplètes. Ces données sont dites censurées. Par exemple, lors de l’étude de l’influence d’un traitement sur la 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.

1. **Objectifs**
2. **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.

1. **Objectifs de délai**

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

1. **Macro-Planning**
2. **Etapes principales du projet**

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

**L’analyse de données se fait en plusieurs étapes :**

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

1. **Dates clé**

* 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 : Soutenance

1. **Ressources & Communication**
2. **Ressources**

* Deux ordinateurs portables sous Windows
* Dépôt Git sur le site internet gitHub.com

1. **Communication**

* WhatsApp
* Facebook
* Travail quotidien à l’ENSIMAG + week end si retard

1. **Risques et dépendances**
2. **Risques majeurs & Actions associées**

|  |  |
| --- | --- |
| 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 |
|  |  |

1. **Dépendances**

* 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 et qui aura pour conséquence directe un retard sur le rendu final.