**ACT-2005TP1 (A remettre le 20 Octobre,2018) dans la chute reservée pour les TP face au secrétariat**

Renseignements généraux pour la partie 1 :

1) Remettre vos réponses rédigées dans un document de WORD. Les résultats importants dans le corps de votre document, mettre dans l’annexe les programmes R utilisés. Ajouter vos commentaires personnels, réflexions. Voir votre TP comme un document préliminaire que vous pourriez rédiger un rapport plus tard et remettre au département R et D de votre compagnie.

2 ) Prendre le temps de le bien faire.

3) Pour la question3, vous pourriez utiliser le output de SAS pour calculer l’estimateur de Kaplan-Meier. Les autres calculs subordonnés pourraient être faits sur R. SAS est beaucoup plus puissant que Excel référé dans votre livre. Au besoin, vous pourriez consulter sous SAS/STATles commandes SAS. L’idée c’est que vous seriez exposé à ce logiciel puissant et au moment propice lorsque vous commencez votre carrière, vous pourriez l’utiliser, ajouter, modifier les procs etc…

Remarque : Pour ce TP l’attention et l’accent sont mis sur les méthodes non paramétriques, la fonction de distribution empirique est la clé pour construire divers estimateurs. Pour le TP2 à venir après la semaine de lecture, l’accent sera sur les méthodes paramétriques.

4 ) Vous annexez vos programmes à la fin du document, pour les simulations par exemple m=50 fois, inclure le premier échantillon obtenu dans l’annexe que vous allez faire référence dans le document. Le document doit être bien relié pour éviter les feuilles qui se détachent avec une page de couverture et le tout dans une enveloppe avec les noms de l’équipe sur l’enveloppe. Les correcteurs vérifieront vos procédures d’une manière plus efficace si vous suivez les instructions.

NB:Vous remettez la version papier dans la chute et il n’ y a pas de remise électronique car c’est plus facile à corriger avec une version papier.

5) Au sujet de la pondération, il y a 10 points accordés pour la présentation générale et les numéros sont avec poids égaux, 30 points par question.

(THÈME : INFÉRENCES NONPARAMÉTRIQUES)

Question 1.(Estimation avec la function de distribution empirique ou repartition empirique)

a)Le coefficient d’asymétrie , est une mesure de symétrie ou d’asymétrie pour la distribution. Si , la distribution est symétrique. Simuler n=100 données d’une loi exponentielle avec une loi exponentielle et voir ces 100 données sont disponibles òu on ne connait pas la loi sous-jacente, estimer avec la fonction de distribution ou répartition empirique ( avec méthode non paramétrique), comparer avec la valeur obtenue avec un histogramme avec les données. Que remarquez vous? Est ce que c’est concordant la valeur de estimé et le graphique de l’histogramme? Vos commentaires, svp.En absence des données réelles,cet échantillon on le voit comme un échantillon des données réelles et dans la pratiqe on n’a qu’un seul échantillon.On va imiter cette situation.

b) On veut non seulement un estimé de , mais aussi un estimé pour la variance de votre estimateur. Pour cela, on va utiliser la méthode ré-échantillonnage , appelée ‘bootstrap’ en anglais .Nous avons besoin de m=50 échantillons tirée de la loi originale qui a donné naissance aux données réelles mais malheureusement on ne connait pas les paramètres de cette loi qui peut etre vue comme une loi dans la famille exponentielle.Ainsi on va utiliser le premier échantillon que vous aviez utilisé dans a) et estimer le paramètre de la loi exponentielle et tirer m=50 échnatillon de cette loi exponentielle estimée et chaque échantillon est de taille n=100;pour chaque échantillon calculer le estimé et on peut utiliser la variance empirique des estimés à travers ces 50 échantillons pour estimer la variance de .On va voir que c’est possible.

Comme la loi exponentielle est sous la forme de densité

, un estimé pour sera et est calculé avec les données de a) vues comme données réelles et vous pouvez utiliser pour tirer m=50 échantillons chacun de tail n=100.

Pour chacun de ces m=50 échantillons vous calculez l’estimateur pour qui est semblable comme dans a) mais calculé avec les données de l’échantillon simulé.La variance de peut etre estimé par une variance empirique comme suit

, la version divisé par m au lieu de m-1 est acceptable aussi car on a m=50. L’estimé de avec échantillon j est dénoté par .La moyenne de ces 50 est dénotée par .

Faire ce travail tel que décrit puis

construire un intervalle de confiance au niveau 0.95 pour .L’estimateur ponctuel de vient du premier échantillon. Voici la formule qui résume

.

On peut faire la remarque que le ré-échantillonnage vous permet d’obtenir qui est la variance estimée pour votre estimateur et celui ci est calculé avec votre échantillon original qui est dans a.

(La percentiles d’une loi t appropriée pourrait etre utilisée au lieu de mais avec peu de différence car m=50)

c) Maintenant, la loi est théorique est révélée, calculer le coefficient d’asymétrie théorique avec la loi exponentielle de moyenne=1, est ce que c’est compatible avec votre estimé ponctuel et par intervalle de confiance obtenus dans b? Vos commentaires, svp.

Remarque : Avec les procédures paramétriques on pourrait faire mieux si l’estimateur non-paramétrique ne performe pas bien mais il faut chercher le modele qui ajuste bien les données.. A venir pour le TP2!

Question 2. (Vous pourriez garder les données de la question 1 et l’énoncé)

a) Estimer avec la distribution empirique. Le tel que c’est défini par le livre aussi appelé `` limted- loss’’ pour les valeurs de

u=,,

,A noter on utilise ici ce qui veut dire la fonction de percentile théorique et non celle qui est empirique.

On suppose que ces u sont des constants car est fixée, fonction de répartition de la loi exponentielle standard avec . En realité on devrait utiliser ,la fonction de répartition empirique avec l’échantillon de a) de la question 1 pour obtenir ces u mais cela entraine des difficultés pour les correcteurs à corriger vos travaux car est aléatoire; alors pour chaque équipe des valeurs légèrement différentes mais malgré tout converge vers ; est définie comme la fonction de répartition d’une loi exponentielle avec Comment varie cette expression ( ) estimée comme fonction de u? Parait- elle logique cette variation? Vos commentaires, svp.

b) Par la méthode bootstrap tel qu’on a discuté dans question 1) avec m=50, construire un intervalle de confiance pour les primes des limted loss pour et u=.Primes signifient espérances.

c) Comparer vos valeurs estimées dans a), les intervalles de confiance obtenus dans b avec les valeurs théoriques des primes du limited -loss provenant de la loi exponentielle standard (. ? Vos commentaires, svp sur les performances de ces estimateurs non-paramétriques.

Question 3( SAS suggéré)

a) Utiliser les 2 colonnes du data set(annexé),utiliser le proc lifetest de SAS approprié pour obtenir l’estimateur de Kaplan-Meier.

b) Faire le graphique de l’estimateur (utiliser SAS), utiliser la formule de Greenwood, calculer un intervalle de confiance au niveau 0.95 pour S (50), méthode linéaire qui est la méthode la plus simple.

c) Re-faire l’intervalle de confiance avec la transformation log (-log) tel que discuté dans livre et dans le cours avec l’énoncé de la question b). Vos commentaires en comparant avec l’intervalle obtenu dans 2). Les calculs pourraient être faits avec R avec un output de SAS.

Annexe :

Voici le jeu de données ( data set )

Temps cens

30 1

40 1

57 1

65 1

65 0

84 1

90 1

92 0

98 1

101 1

Remarques et commentaires :

Légende :

Cens=1, donnée complète, c’est-à-dire l’âge de décès est enregistrée. Cens=0, donnée incomplète par exemple âge= 65, cens=0 veut dire l’âge de décès n’est pas enregistré .On sait seulement que la personne a survécu l’âge 65 et on perd l’information après l’âge 65.Vous pourriez constater que en assurance vie on rencontre souvent ces types de données mais pas aussi courantes pour les réclamations en IARD.

Suggestion : R est capable de faire des intégrations, la commande integrate, ou quad dans le package pracma à utiliser au besoin pour vérifier vos calculs.

Commandes SAS pour trouver l’estimateur Kaplan-Meier avec proc lifetest

Data NOM;

Input temps, cens;

Datalines:

(Rentrer les données)

;

Run

Proc lifetest data=NOM;

Temps\*cens(0);

Run;