Université Sorbonne Paris Nord, Institut Galilée
Année universitaire 2020–2021
mallein@math.univ-paris13.fr

TP 3 – Analyse de données

1 Estimateur, intervalle de confiance d'un n-échantillon

Exercice 1 (Loi normale)**.**

1. Simulez un échantillon de 100 gaussiennes centrées réduites (fonction `rnorm`).
2. Calculez pour cet échantillon la moyenne et la variance empirique.
3. Représentez graphiquement cet échantillon à l'aide d'un histogramme.
4. On suppose que l'on a oublié la valeur de l'espérance, mais qu'on sait que la variance est égale à 1. Déterminer un invervalle de confiance à 95% pour l'espérance.
5. Recommencez 1000 fois la construction de cet intervalle de confiance sur des échantillons indépendants. Combien de ces intervalles de confiance contiennent 0 ?

Exercice 2 (Larves d'éphémères)**.** Le tableau suivant donne la distribution des longueurs de larves d'éphémères. Donnez un intervalle de confiance à 95% pour leur longueur moyenne.

Table 1 – Longueurs à la naissance en millimètres de 64 larves d'éphémères

longueur X (en mm)	Nombre de larves
$11 \leq X < 15$	2
$15 \leq X < 17$	4
$17 \leq X < 19$	7
$19 \leq X < 21$	9
$21 \leq X < 23$	6
$23 \leq X < 25$	15
$25 \leq X < 27$	12
$27 \leq X < 29$	5
$29 \leq X < 33$	4

Exercice 3 (Lait contaminé)**.** Pour savoir si un traitement particulier permet de réduire le nombre de bactéries dans le lait écrémé, on a compté le nombre de bactéries avant et après traitement dans des échantillons de lait écrémé. Les résultats sont exprimés en logarithme du nombre de bactéries dénombrées.

Table 2 – Logarithme du nombre de bactéries avant et après traitement dans 12 échantillons de lait écrémé.

Échantillon	Avant traitement	Après traitement
1	6.98	6.95
2	7.08	6.94
3	8.34	7.17
4	5.30	5.15
5	6.26	6.28
6	6.77	6.81
7	7.03	6.59
8	5.56	534
9	5.97	5.98
10	6.64	6.51
11	7.03	6.84
12	7.69	6.99

1. Représentez graphiquement les données par un nuage de points.
2. Peut-on supposer l'indépendance des données avant et après traitement ?
3. Le traitement semble-t-il diminuer le nombre de bactéries. Construisez un intervalle de confiance à 95% pour justifier votre réponse.

2 Maximum de vraisemblance d'un modèle

Exercice 4 (Survie de patients leucémiques)**.** On donne dans le tableau 1 ci-dessous le temps de survie y (en mois) en fonction du nombre de globules blancs initial x, pour 17 patients atteints de leucémie.

1. Représentez graphiquement y en fonction de x. Les données présentent-elles une tendance ?
2. Que pensez-vous de l'utilisation d'une régression linéaire (obtenue grâce à `lm`) sur ces données ?
3. Nous réalisons l'hypothèse statistique que les variables Y_i suivent des lois exponentielles indépendantes, avec pour espérance
$$\mathbf{E}(Y_i) = \exp(\beta_1 + \beta_2 x_i).$$
Donnez une formule permettant de calculer la vraisemblance de ces observations. Définissez une fonction `logvrais` prenant en argument un vecteur de longueur 2 permettant le calcul de la log-vraisemblance en (β_1, β_2). Essayez de maximiser cette fonction à l'aide de `nlm` à partir de différentes initialisations. Quel est le résultat obtenu ?
4. Tracez une représentation 3D (grâce à `persp`) de la surface de log-vraisemblance pour $\beta = (\beta_1, \beta_2) \in [4, 5] \times [-0.01, 0.01]$. Qu'observez-vous ?
5. Donnez l'équation définissant l'estimateur du maximum de vraisemblance du couple (β_1, β_2) dans le modèle défini. Calculer à nouveau le maximum de vraisemblance en utilisant `nlm`. Quele est la valeur du Hessien au maximum ? Donner une estimation de l'information de Fisher.
6. Tracez le graphes des observés et des ajustés (courbe estimée superposée aux points observés).

TABLE 3 – Temps de survie y en mois en fonction du log10 du nombre de globules blancs initial x pour 17 patients atteints de leucémie.

x	65	156	100	134	16	108	121	4	39	143	56	26	22	1	1	5	65
y	29	18	38	30	44	56	55	69	42	47	53	91	94	148	148	112	148

Exercice 5 (Jeu de données simulées)**.**

1. Construisez une fonction `modeleMelange` prenant en entrée un vecteur de longueur 5 $(p, \mu_1, \sigma_1, \mu_2, \sigma_2)$ et renvoyant une variable aléatoire de loi $\mathcal{N}(\mu_1, \sigma_1^2)$ avec probabilité p et de loi $\mathcal{N}(\mu_2, \sigma_2^2)$.
2. Générez un vecteur `observations` de longueur 1000 constitué de réalisations de `modeleMelange` pris avec les paramètres $p = 0.3$, $\mu_1 = -5$, $\sigma_1 = 2$, $\mu_2 = 3$ $\sigma_2 = 1$.
3. Tracez l'histogramme de `observations`. Un modèle de mélange gaussien se justifie-t-il au vu des données ?
4. Construisez une fonction `logVrais` permettant de calculer la log-vraisemblance de l'échantillon.
5. Déterminez le maximum de vraisemblance grâce à `nlm`. Essayez différentes initialisations. Pour quelles initialisations le programme termine-t-il ?
6. Faites à nouveau les questions 2 à 5 avec les paramètres $p = 0.9$, $\mu_1 = \mu_2 = 0$, $\sigma_1 = 1$, $\sigma_2 = 10$.

3 Importation de données, modèle linéaire gaussien

Exercice 6 (Survie de patientes atteintes de cancer du sein)**.** On étudie la durée de survie Y de femmes atteintes de cancer du sein soumises à trois traitements, A, B et C. Ces durées figurent dans le ficher joint `survieS.data`.

1. Lisez toutes les données avec `read.table`. Créez un `data.frame` contenant un vecteur des temps de survie, un vecteur des âges et un vecteur des traitements.
2. Étudiez les temps de survie relatifs aux traitements B et C. Calculer leurs moyennes et leurs variances. Représentez les données.
3. Peut-on considérer que les traitements sont équivalents ? Vous expliquerez les choix faits dans la mise en place de votre test, ainsi que les résultats de `t.test`.
4. Sans tenir compte de l'âge d'apparition du cancer, tester l'existence d'un effet traitement (on utilisera les fonctions `lm` et `anova`, éventuellement `qf`).
5. On soupçonne l'existence d'un lien entre l'âge d'apparition du cancer et la durée de survie. Pour chaque traitement, représentez la durée de survie en fonction de l'âge. Superposez un lissage obtenu à partir de la fonction `lowless`. Quel modèle permet de ces données ?

6. On note X l'âge d'apparition du cancer, et on envisage le modèle général $M1$ suivant

$$Y = \begin{cases} a_A + b_A X + \epsilon & \text{si traitement } A \\ a_B + b_B X + \epsilon & \text{si traitement } B \\ a_C + b_C X + \epsilon & \text{si traitement } C \end{cases}$$

Estimez les paramètres du modèle à l'aide de la fonction `lm`. Testez le sous-modèle $M0$ correspondant à l'égalité des droites $Y = a + bX + \epsilon$. Que conclue-t-on quant à un effet traitement ?