Travaux Dirigés ISV51 - Statistiques descriptives

Julien Chiquet

23 octobre et 6 novembre 2015

Objectifs de la séance

- lire/écrire un tableau de données dans un fichier
- résumés numérique élémentaire
- bases de l'analyse descriptive
- représentations graphiques élémentaires

Exercice 1: lecture/écriture d'un tableau de données - chromosomes

1. Créer un tableau à 24 lignes et 3 colonnes en lisant le fichier chromosomes.txt avec la fonction `read.table`. Chaque ligne représentera un chromosome humain (22 autosome, 2 chromosomes sexuels) et les colonnes seront respectivement leur noms, nombre de gènes, et longueur en bases.
2. Représenter Le nombre de gène en fonction du nombre de bases.
3. Ajouter une colonne supplémentaire au tableau qui spécifie pour chaque chromosome s'il est autosome ou pas.
4. Calculer le nombre total de paire de base d'un génome humain (pour un homme, puis une femme).

Exercice 2: lecture données, graphiques - somnifère

Pour étudier l'effet d'un somnifère, on mesure chez 20 patients le nombre d'heures de sommeil supplémentaires par rapport à la durée moyenne de leur nuit sans traitement. On obtient les résultats suivants:

# patient	extra
1	0.7
2	-1.6
3	-0.2
4	-1.2
5	-0.1
6	3.4
7	3.7
8	0.8
9	0.0
10	2.0
11	1.9
12	0.8
13	1.1
14	0.1
15	-0.1
16	4.4
17	5.5
18	1.6
19	4.6
20	3.4

1. Saisir ces données dans un vecteur.
2. Faites un résumé numérique.
3. Tracer un diagramme en tige et feuille.
4. Tracer la fonction de répartition empirique puis l'histogramme normalisé des données dans la même fenêtre graphique.
5. Ces données sont en fait issues de deux groupes d'individus: apposer une variable indiquant le groupe associé à l'observation de la variable `extra` sachant que les 10 premiers individus sont issus du groupe 1 et les 10 suivants du groupe 2 (utiliser, par exemple, la commande `data.frame`). Faire un résumé statistique pour chaque groupe et tracer alors les boîtes à moustaches des observations selon les groupes. Qu'en pensez-vous ?

Exercice 3: variables qualitative - variations génétiques dans les populations humaines.

1. Charger le jeu de données `hdpg` du package `ade4` et lire son descriptif.
2. Nous considérerons le tableau `hdpg\$ind` qui décrit l'échantillon des 1066 individus de l'étude.
3. Combien de populations différentes participent à l'étude ?
4. Dresser les tableaux des effectifs des variable population, région et sexe.
5. Transformer ces tableaux en tableaux de fréquences.
6. Représenter vos tableaux de fréquence par des diagramme en bâton, et par des camemberts.
7. Représenter les fréquences cumulées.
8. Commenter les représentations.

Exercice 4: variable quantitative - coefficient de Gini

Le coefficient de Gini permet de mesurer l'inégalité des revenus dans une population. Si tous les individus gagnent le même salaire le coefficient de Gini vaut 0 (situation égalitaire), alors que si un seul individu gagne tous le revenu disponible et les autres rien l'index de gini vaut 1. Les états-unis ont par exemple un coefficient de Gini de 0.47.

1. Charger le jeu de données gini.Rdata.
2. Sélectionner les lignes du tableau correspondant à l'année 2007.
3. Tracer un diagramme en tige et feuille des coefficients.
4. Tracer l'histogramme des coefficients.
5. Tracer l'histogramme lissé des coefficients.
6. Tracer le boxplot des coefficients.
7. Tracer un diagramme des fréquences cumulées des coefficients.
8. Écrire une fonction R qui rende les pays de coefficient Gini d'index maximum et minimum.
9. Classer les pays par leur coefficient de Gini.
10. Calculer la moyenne, la variance, le coefficient d'asymétrie, le coefficient d'aplatissement pour la distribution des coefficients de gini. Commenter.

11. Combien de pays sont plus égalitaires que la France en europe.

Exercice 5: simulation et graphiques - puces à ADN

Les distributions des intensités moyennes des spots d'ADNc correspondants aux gènes exprimés et non exprimés peuvent être modélisées par deux gaussiennes. On suppose que la première distribution a une espérance $\mu^{\mathrm{e}} = 1000$ et un écart type $\sigma^{\mathrm{e}} = 100$; la seconde distribution a pour espérance $\mu^{\mathrm{ne}} = 400$ et pour écart type $\sigma^{\mathrm{ne}} = 150$.

Chaque gène correspond à 4 spots répliqués. L'expression d'un gène est définie comme la moyenne des 4 spots qui lui sont associés.

Analyse élémentaire du modèle

1. Créer sous `R` les variables `mu.e`, `sigma.e`, `mu.ne` et `sigma.ne` et affectez-y les valeurs de l'énoncé.
2. On note S^{e} la variable aléatoire décrivant l'intensité d'un spot correspondant à un gène exprimé. Quelle est la probabilité pour que S^{e} ait une valeur inférieure ou égale à 700 ?
3. On note G^{e} la variable aléatoire décrivant le niveau d'expression d'un gène exprimé. Quelle est la probabilité pour que G^{e} ait une expression inférieure ou égale à 700 ?
4. On introduit la variable aléatoire G^{ne} pour les gènes non exprimés. Quelle est la valeur seuil t telle que la probabilité d'avoir G^{e} inférieure ou égale à t soit égale à la probabilité d'avoir G^{ne} supérieure à t ?
5. Quelle est la probabilité d'avoir un gène exprimé dont l'expression est inférieur à t (faux négatif) ?
6. Quelle est la probabilité d'avoir un gène non exprimé dont l'expression est supérieure à t (faux positif) ?

Simulations et graphiques

1. Générer $n = 1000$ intensités de spots correspondant aux gènes exprimés et non exprimés. Les stocker dans les vecteurs `spots.e` et `spots.ne`. Parmi tous les spots générés, stocker la plus petite et la plus grande valeur observée dans des variables `MIN` et `MAX`.
2. Créer deux objets de classe histogramme, sans les tracer, correspondant à chacune des deux populations de spots et stocker les dans des variables `hist.e` et `hist.ne`.
3. Tracer sur un même graphique les deux histogrammes normalisés et les densités théoriques (fonction `curve`). Utiliser deux couleurs différentes pour les deux populations de spots. Apposer une légende au graphe (commande `legend`).
4. Tracer sur un même graphique les densités théoriques des gènes exprimés et non exprimés. Faire une légende. Puis, à l'aide de la commande `polygon`, représenter l'aire sous courbe correspondant à la probabilité pour qu'un gène non exprimé ait une expression inférieure à 300. Enfin, tracer une droite verticale indiquant l'emplacement du seuil t (commande `abline`).