1 Test d'ajustement à l'espérance d'une loi normale

Successivement et pour deux jeux de données différents (avec un minimum de 4 variables... par exemple : *seeds* du site *UCI repository* ; *crabs* du package *MASS* de *R* ; *USArrests*, *state.x77* ou *wine* disponibles dans les jeux de données de *R*...), préparer une présentation des données incluant au moins :

— Quelques éléments sur la nature, les valeurs, la position et la dispersion de chaque variable. On pourra inclure des graphiques.

— Quelques représentations graphiques rendant compte de la structure multidimensionnelle des données lorsqu'on considère les projections des variables quantitatives sur les droites, plan, sous-espaces de dimension 3, etc. définis par la base canonique.

— Une ACP des variables quantitatives et son analyse.

On pourra utiliser les outils de base de *R* mais aussi, par exemple, le package *ggplot2* pour les représentations graphiques et pour le package *FactoMineR* pour l'ACP.

2 Controverse sur la théorie de l'hérédité de Mendel

En 1939, une scientifique soviétique, Ermolaeva, publiait les résultats d'une étude reproduisant des expériences analogues à celles par lesquelles Mendel avait abouti à sa théorie de l'hérédité au XIX^{e} siècle. Ermaloeva (et son équipe) a étudié les caractères de graines issues de l'auto-pollinisation de plantes hétérozygotes portant un allèle dominant et un allèle récessif pour un gène donné (ces plantes étant elles-mêmes obtenues par le croisement d'une plantes homozygote pour l'allèle dominant et d'une plante homozygote pour l'allèle récessif). Si la théorie de Mendel était vraie (ce qui ne fait plus de doute aujourd'hui), on devait observer $\frac{3}{4}$ de graines au caractère dominant et $\frac{1}{4}$ de graines au caractère récessif... à l'aléas près !

Ermolaeva a prétendu que ses données infirmaient la théorie de Mendel. Kolmogorov (le célèbre mathématicien, soviétique également) s'est penché sur les mêmes données et a affirmé le contraire...

Nous allons à notre tour nous pencher sur les données que Ermolaeva rapporte pour deux de ses expériences et qui nous sont fournies par Stark and Seneta (2011). Attention : nous ne prétendons certainement pas pouvoir trancher la question à l'occasion de ce TP. D'une part, il faudrait étudier très sérieusement les articles des différents protagonistes : les conditions d'expérience, la façon dont les données ont été produites, collectées, traitées. Par exemple, Stark and Seneta (2011) discute de certaines incohérence dans la présentation des données par Ermolaeva : pour se faire un avis sur ces éventuelles incohérences, il faudrait prendre le temps de se pencher dessus (ce que je n'ai pas fait et ce pour quoi le matériel fourni ici ne suffirait absolument pas). D'autre part, il faudrait alors réfléchir à la meilleure approche statistique pour traiter ces données et sans doute utiliser des méthodes plus élaborées que celles que nous aurons le temps d'aborder à l'occasion d'un TP. Même alors, rien ne garantit que les méthodes statistiques permettent de trancher : une conclusion possible serait par exemple que les expériences ne sont pas suffisamment nombreuses pour conclure.

Anecdote : Lorsque Fisher a étudié les célèbres données de Mendel lui-même, il a trouvé, non seulement que celles-ci validaient sa théorie... mais même qu'elles la validaient trop bien : la probabilité d'une aussi bonne adéquation entre les données et la théorie était trop grande ! Fisher (Fisher, 1936, Section 4, p. 132) suppose qu'un assistant de Mendel a dû arranger quelque peu les données dans le sens attendu...

La première expérience porte sur la couleur de l'enveloppe des graines, qui peuvent être gris-brun (dominant) ou blanc (récessif). Les résultats retenus (correspondant à la table 5 de Stark and Seneta, 2011) sont dans le fichier *SeedCoat.csv*. La deuxième porte sur la couleur du cotylédon : jaune (dominant) ou vert (récessif) et les données correspondantes (table 6 de Stark and Seneta, 2011) sont dans le fichier *Cotyledon.csv*. On traitera ces deux jeux de données indépendamment mais en suivant les mêmes étapes décrites ci-dessous.

Récupérez les données. Regroupez-les selon la variable *Set* (qui décrit des ensembles de graines correspondant à des plantes différentes mais issues du même croisement et que nous choisissons de regrouper, ce qui est discutable). Calculez ensuite la p-valeur des tests suivants et concluez, en choisissant soigneusement qui sont les variables dans chaque cas et en discutant notamment des hypothèses sous-jacentes à chaque fois :

— Le test d'ajustement au paramètre d'une loi binomiale sans approximation normale de la binomiale.

— Le test d'ajustement au paramètre d'une loi binomiale avec approximation normale de la binomiale.
— Le test de Kolmogorov pour tester qu'une statistique bien choisie suit une loi normale $\mathcal{N}(0,1)$.
— Un test de χ^2 sur les proportions (parmi toutes les graines) de caractères dominants et récessifs.

2.1 Le test de Kolmogorov

C'est le test que Kolmogorov a utilisé sur ce jeu de données et qui porte aujourd'hui son nom. Je ne sais pas s'il l'a développé à cette occasion cependant, ni même s'il l'avait déjà présenté avant.

Il repose sur un résultat remarquable : soit F_0 une fonction de répartition continue et F_n la fonction de répartition empirique associée (c'est-à-dire la fonction de répartition de la loi empirique associée à un échantillon $X_1, \dots, X_n$ iid de la loi F_0 : $F_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i \leq x}$: notez que c'est une statistique, à valeurs dans l'ensemble des fonctions cadlag). Alors la loi de $T_n = \sqrt{n}\sup_{x\in\mathbb{R}} |F_n(x) - F_0(x)|$ ne dépend pas de F_0 ! Kolmogorov a même montré que

$$\forall \lambda > 0, \mathbb{P}\left(T_n > \lambda\right) \xrightarrow[n\to\infty]{} 2\sum_{k=1}^{\infty} (-1)^{k+1} \mathrm{e}^{-2k^2\lambda^2}.$$

Notez (et démontrez !) que

$$\begin{aligned}
&\forall x \in \mathbb{R},\\
&\sup_{x\in\mathbb{R}} |F_n(x) - F_0(x)|\\
&\qquad = \sup\Big\{ \frac{i}{n} - F_0(X_{(i)}) : i \in \{1,\dots,n\}\Big\} \cup \Big\{ F_0(X_{(i+1)}) - \frac{i}{n} : i \in \{1,\dots,n-1\}\Big\} \cup \big\{F_0(X_{(1)})\big\},
\end{aligned}$$

où $X_{(1)}, \dots, X_{(n)}$ est l'échantillon $X_1, \dots, X_n$ ordonné : $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. Cette remarque est très utile pour la mise en œuvre de la méthode.

Voir par exemple Rivoirard and Stoltz (2012, section 8.2), Bickel and Doksum (2015, Example 4.1.5) ou Lehmann and Romano (2006, Section 14.2).

2.2 Le test du χ^2

On peut utiliser dans ce TP un exemple de test du χ^2. Il faut pour cela connaître la loi qui porte ce nom : la loi du $\chi^2(k)$ (« à k degrés de liberté ») est la loi de $Z_1^2 + \cdots + Z_k^2$ lorsque $Z_1, \dots, Z_k$ est un échantillon iid de la loi $\mathcal{N}(0,1)$. Elle est donc d'espérance k et de variance $2k$. Il s'agit de la loi $\Gamma(\frac{k}{2}, 2)$.

Il faut ensuite savoir que, si $(N_1, \dots, N_k)$ suit une loi multinomiale $\mathcal{M}(n, p_1, \dots, p_k)$ où $n \in \mathbb{N}^*$ et $p_1, \dots, p_k$ sont des réels positifs tels que $p_1 + \cdots + p_k = 1$ (i.e. $\mathbb{P}\left(N_1 = n_1, \dots, N_k = n_k\right) = \frac{n!}{n_1!\dots n_k!} p_1^{n_1} \dots p_k^{n_k} \mathbb{1}_{n_1+\cdots+n_k=n}$: N_i est le nombre de fois où on a observé i en tirant n variables aléatoires indépendantes dont la probabilité de voir j est p_j pour tout $j \in \{1,\dots,k\}$), alors $\sum_{i=1}^k \frac{(N_i - np_i)^2}{np_i} \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(k-1)$. On considère l'approximation $\sum_{i=1}^k \frac{(N_i - np_i)^2}{np_i} \sim \chi^2(k-1)$ acceptable lorsque $\forall i \in \{1,\dots,k\}, np_i \geq 5$.

Voir par exemple Rivoirard and Stoltz (2012, section 6) ou Lehmann and Romano (2006, Section 14.3).

Références

Bickel, P. J. and Doksum, K. A. (2015). *Mathematical statistics : basic ideas and selected topics*, volume 1. CRC Press.

Fisher, R. A. (1936). Has mendel's work been rediscovered ? *Annals of science*, 1(2) :115–137.

Lehmann, E. L. and Romano, J. P. (2006). *Testing statistical hypotheses*. Springer.

Rivoirard, V. and Stoltz, G. (2012). *Statistique mathématique en action*. Vuibert.

Stark, A. and Seneta, E. (2011). An kolmogorov's defence of mendelism. *Genetics and molecular biology*, 34(2) :177–186.

The End.