M2 Mycologie : Notes et compléments

Inférence et test d'hypothèse

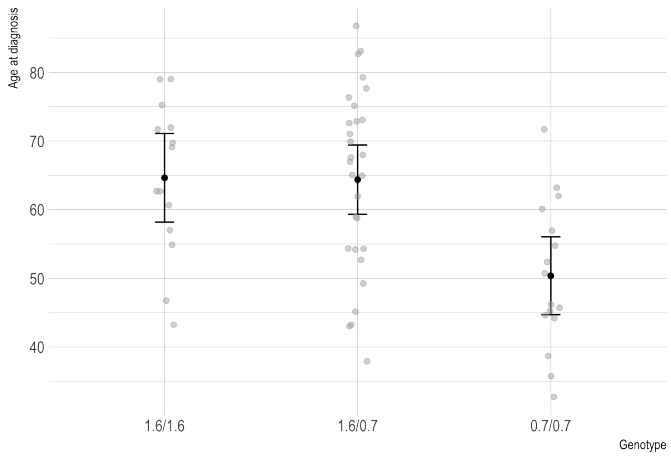
Comme dans un essai clinique, on cherche donc à comparer l'effet d'un traitement A à celui d'un traitement B de référence (ou placebo), qui sert de comparateur. La différence observée entre les effets de A et B peut faire l'objet d'un test statistique. Celui-ci permet de confronter la valeur observée à celles pouvant résulter de simples fluctuations d'échantillonnage (c.a.d. des différences dûes au "hasard", signifiant ici l'absence d'une réelle différence).

D'un autre côté, il existe un risque β de ne pas être en mesure de rejeter l'hypothèse nulle lorsqu'une réelle différence existe. Le complément $1 - \beta$, appelée puissance du test, représente donc la probabilité de rejeter correctement l'hypothèse nulle en faveur de l'hypothèse alternative.

Analyse de variance

On utilise des données collectées dans le cadre d'une étude sur le polymorphisme du gène du récepteur estrogène (3 niveaux) en fonction de l'âge de diagnostic des individus [1].

		N	Mean	SD
genotype	1.6/1.6	14	64.6429	11.1811
	1.6/0.7	29	64.3793	13.2595
	0.7/0.7	16	50.3750	10.6388
Overall		59	60.6441	13.4943



Soit y_{ij} la j$^{\text{ème}}$ observation dans le groupe i. Le modèle d'ANOVA ou "effect model" s'écrit

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

où μ désigne la moyenne générale, α_i l'effet du groupe i, et $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ un terme d'erreur aléatoire. On impose généralement que $\sum_{i=1}^{k} \alpha_i = 0$.

L'hypothèse nulle se lit $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k$. Sous cette hypothèse d'égalité des moyennes de groupe, la variance entre groupes (”between”) et la variance propre à chaque groupe (”within”) permettent d'estimer σ^2. D'où le test F d'égalité de ces deux variances. Sous H_0, le rapport entre les carrés moyens inter et intra-groupe (qui estiment les variances ci-dessus) suit une loi F de Fisher-Snedecor à $k-1$ et $N-k$ degrés de liberté.

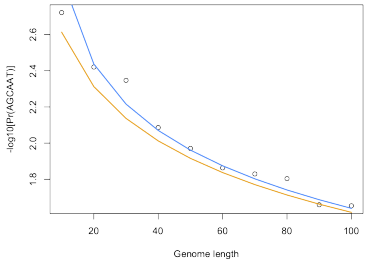
Voici les résultats produits à l'aide du logiciel R :

```
            Df Sum Sq Mean Sq F value  Pr(>F)
genotype     2   2316    1158    7.86 0.00098
Residuals   56   8246     147

 Pairwise comparisons using t tests with pooled SD

data:  age and genotype

        1.6/1.6 1.6/0.7
1.6/0.7 0.947   -
0.7/0.7 0.002   5e-04

P value adjustment method: none
```

Recherche de motifs

Let us consider a sequence of size k and a window of size n (n much smaller than the genome length). If all nucleotides are assumed independant, then the probability of not finding this particular sequence is $(1 - 1/4^k)$ (using the complementary is useful to avoid dealing with the number of times this sequence can appear) times the total number of positions available in the window (n or n - k + 1, but if $k \ll n$ this correction factor can be omitted). Hence the answer : $1 - (1 - 1/4^k)^n$. Pattern Markov chains would provide a more precise answer in this case.

La figure ci-contre représente les résultats d'une simulation avec les estimations ponctuelles (points de couleur noire) et les valeurs attendues avec (bleu) ou sans (orange) correction pour la longueur de la séquence. Cette dernière n'est vraiment influente que dans le cas des petites tailles de fenêtre.



Seconde loi de Mendel

On calcule dans un premier temps la probabilité d'observer n Aa Bb descendants après k générations. Il s'agit d'une succession d'événements de Bernoulli jusqu'à la génération k, d'où $P(b,k) = \binom{2^k}{n}(1/4)^n(1-1/4)^{2^k-n}$, qui n'est autre que la fonction de densité d'une variable aléatoire suivant une loi Binomiale de paramètre $n = 2k$ (il y a deux enfants à chaque étape) et $p = 1/4$. Ensuite il s'agit de trouver la probabilité qu'au moins N organismes se trouvent à la génération k, qui vaut $1 - \sum_{x=0}^{N-1} P(X = x)$, où X est une variable aléatoire représentant le nombre de descendants, c'est-à-dire 1 moins tous les cas $P(X \neq N)$.

Le terme $1/4$ vient du fait que la probabilité d'observer un descendant de ce sous-type est uniforme quelque soit le croisement considéré.

Prolifération bactérienne

Le nombre moyen de bactéries par boite est 5. On suppose que le nombre de colonies par boite est le même que le nombre moyen de bactéries par boite. On a alors $P(X = 0) = \frac{5^0 e^{-5}}{0!} = 0.0067$, soit approximativement 0,67 % de chance.

On remarquera que la probabilité qu'il y ait au moins une colonie sur la boite de Pétri vaut $P(X > 0) = 1 - P(X = 0) = 1 - 0.0067 = 0.9933$ (99,3 % de chance) !

Le modèle de Jukes-Cantor en phylogénie

On souhaite comparer deux espèces (eucaryotes) ayant un ancêtre commun. Des mutations aléatoires intervenant sur le génome au cours de leur évolution, un alignement des deux séquences ne donnera pas 100 % d'identité.

Matrice de transition (ACGT) :

$$\begin{pmatrix} -0.886 & 0.190 & 0.633 & 0.063 \\ 0.253 & -0.696 & 0.127 & 0.316 \\ 1.266 & 0.190 & -1.519 & 0.063 \\ 0.253 & 0.949 & 0.127 & -1.329 \end{pmatrix}$$

Si l'on se trouve dans l'état `A`, on y restera un temps exponentiel de paramètre $-q_{ii} = 0.886$. La probabilité d'observer la transition `A->C` est donnée par $-q_{ij}/q_{ii} = \frac{0.190}{0.886}$.

En somme, si l'on part de `A`, on attendra $\exp(0.886) = 2.425$ unités de temps, et la probabilité de transitionner vers un autre état sera de 0.214 (pour C), 0.714 (G) et 0.071 (T). La transition la plus probable `A->G` implique d'attendre $\exp(-0.633) = 0.531$ unité de temps.

PCR et processus de branchement

From a practical point of view, it is often desired to know N_0 in advance, although N_0 cannot be identified uniquely. Note, however, that

it is possible to consistently estimate the mathematical expectation of the reproduction m with th estimator $\hat{m}_n = \sum_{k=0}^{\infty} k\hat{p}_{k_n} = \frac{Y_{(n+1)}-1}{Y_n}$, that is m can be estimated from the observed size of each generation. Hence, the following estimator has been proposed : $\hat{N}_{0n} = \frac{Z_n}{\hat{m}_n}$. As an illustration, if the success rate of the PCR is 80% (i.e., $p = 0.8$), its variance equals $\frac{1-p}{1+p} = 0.11$.

Références

[1] W. D. DUPONT. *Statistical Modeling for Biomedical Researchers*. 2nd. Cambridge University Press, 2009.