Concepts de probabilités

Probabilités et statistique pour la biologie (STAT1)

Jacques van Helden

2018-11-09

Contents

Définitions de la probabilité

Définition fréquentielle de la probabilité

Lors d'une expérience aléatoire, la ***probabilité*** d'un événement A est la limite de sa fréqunce de réalisation quand le nombre d'essais tend vers l'infini.

$$\mathrm{P}(A) = \lim_{n \to \infty} \frac{n_A}{n}$$

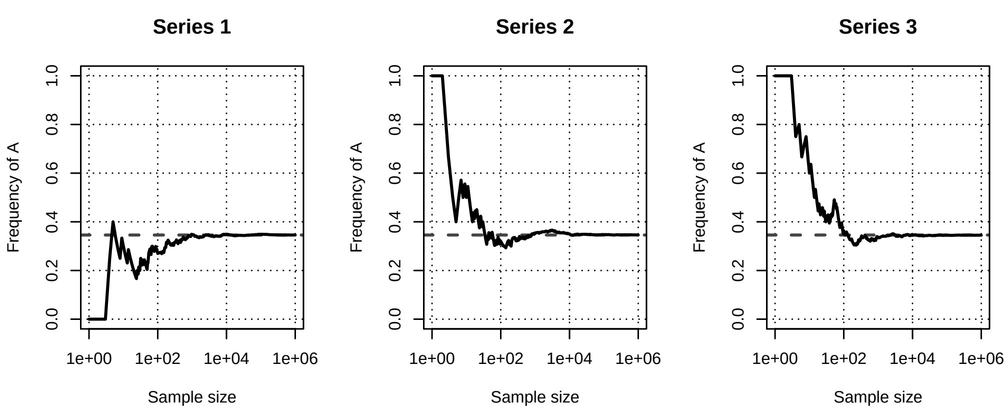

Figure 1: Fréquences de A lors du tirage (avec remise) de positions aléatoires dans le génome de *Mycoplasma genitalium*.

Probabilités pour des ensemble finins

Supposons qu'on tire des éléments dans un ensemble fini, en considérant certains tirages comme des succès (A). Dans cette situation, la probabilté est définie comme le rapport entre le nombre de tirages pouvant être considérés comme des succès (n_A) et le nombre total de tirages possibles (n).

$$\mathrm{P}(A) = \frac{n_A}{n}$$

Exemple: sélection aléatoire de 4 cartes dans un jeu de 52 cartes. Quelle est la probabilité d'avoir un carré (4 cartes identiques) ?

Le jeu de carte comporte 13 valeurs (As, 2, 3, ..., Dame, Roi), il y a donc 13 possibilités d'obtenir un carré: $n_A = 13$.

Le nombre total de tirages de 4 cartes parmi 52 est fourni par le coefficient binomial:

$$n = \binom{52}{4} = 270725$$

$$\mathrm{P}(A) = \frac{n_A}{n} = \frac{13}{\binom{52}{4}} = \frac{13}{270725} = 4.8 \times 10^{-5}$$

Exercice: tirage de tétranucléotides dans un génome

On tire aléatoirement une position génomique. Quelle est la probabilité pour que le tétranucléotide commençant à cette position corresponde aux critères suivants ?

a. Etre composé uniquement de A.
b. Etre composé de 4 résidus distincts.
c. Ne comporter aucun A.

Formulez explicitement le raisonnement qui vous amène à la formule de calcul. Indiquez ensuite la formule générale (avec des symboles), puis la formule particulière avec les valeurs numériques. Il n'est pas nécessaire de calculer le résultat final.

Probabilités d'événements combinés

Exclusion mutuelle

On désigne des événements de **mutuellement exclusifs** quand la réalisation de l'un rend impossible la réalisation des autres. Leur **probabilité jointe** (A_1 et A_2) est donc nulle.

$$A_1, A_2 \text{mutuellement exclusifs} \iff \mathrm{P}(A_1 \wedge A_2) = 0$$

Le symbole $\wedge$ correspond au ***et*** logique.

Si deux événements A_1 et A_2 sont mutuellement exclusifs, la probabilité de réalisation de l'un ou de l'autre est la somme de leurs probabilités.

$$\mathrm{P}(A_1 \wedge A_2) = 0 \iff \mathrm{P}(A_1 \vee A_2) = \mathrm{P}(A_1) + \mathrm{P}(A_2)$$

Complémentarité

Un ensemble d'événements $A_1, A_2, \ldots, A_m$ sont dit **complémentaires** s'ils sont mutuellement exclusifs et exhaustifs (il n'existe pas d'événement possible en dehors de l'ensemble).

La **probabilité de l'union** d'événements complémentaires vaut 1.

$$A_1, A_2, \ldots A_m \text{complementary} \Rightarrow \mathrm{P}(A_1 \vee A_2 \vee \ldots \vee A_m) = \mathrm{P}(A_1) + \mathrm{P}(A_1) + \ldots + \mathrm{P}(A_m) = 1$$

Indépendance stochastique

Deux événements sont **stochastiquement indépendants** si la réalisation de l'un n'affecte pas la probabilité de réalisation de l'autre.

La **probabilité jointe** d'une série d'événements indépendants est le produit de leurs probabilités.

$$A_1, A_2, \ldots A_m \text{ independent } \Rightarrow \mathrm{P}(A_1 \wedge A_2 \wedge \ldots \wedge A_m) = \mathrm{P}(A_1)\,\mathrm{P}(A_2) \ldots \mathrm{P}(A_m)$$

Formule générale de probabilités combinées

$$\mathrm{P}(A \vee B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \wedge B)$$

Schéma de Bernoulli

Un ***essai de Bernoulli*** est une expérience aléatoire qui peut résulter en deux événements possibles, dénommés ***succès*** et ***échec***, chacun étant associé à une probabilité.

Un ***schéma de Bernoulli*** est une série d'essais qui satisfont les conditions suivantes:

1. Indépendance: le résultat d'un essai n'affecte pas les probabilités de succès de l'essai suivant.
2. La probabilité de succès est identique pour chacun des essais.

Schéma de Bernoulli généralisé

On peut généraliser la définition précédente en considérant une série d'essais pouvant résulter en un nombre finis de résultats possibles, associés chacun à une probabilité. Si les essais successifs sont indépendants et les probabilités des événements constantes au fil des essais, on parle de ***schéma de Bernoulli généralisé***.

Exemple: modèle de probabilité des nuclétides dans le génome de la levure, basé sur les fréquences nucléotidiques.

Résidu	Occurrences génomiques	Fréquence génomique
A	3766191	0.3098064564636
C	2320522	0.1908858838986
G	2316991	0.1905954242278
T	3752889	0.3087122354100

Dans le cadre d'un tirage aléatoire, on va considèrer que chaque résidu a une probabilité particulière: $\mathrm{P}(A) = 0.31$, $\mathrm{P}(C) = 0.19$, $\mathrm{P}(G) = 0.19$, $\mathrm{P}(T) = 0.31$.

Les modèles de Bernoulli conviennent-il pour les séquences biologiques ?

Le schéma de Bernoulli présente un intérêt évident: sa facilité d'utilisation. Cependant il n'est pas très approprié pour modéliser les successions de résidus dans les séquences macromoléculaires, pour plusieurs raisons.

- Dans les génomes, les fréquences de nucléoides varient en fonction du contexte (codant, non-codant, ...).
- Dans les séquences codantes, dépendance entre les résidus d'un même codon, notamment au niveau de la dégénérescence du 3ème résidu.
- Dans les régions non-codantes, les oligonucléotides poly-A et poly-T sont plus fréquents (propriété agrégative).
- Dans les séquences protéines, les contraintes structurelles ont favorisé (par sélection) certaines successions d'acides aminés et défavorisé d'autres.

Les ***modèles de Markov*** permettent une modélisation plus fine des séquences biologiques, en exprimant la dépendance entre résidus voisins.

Exercices

1. On tire un nucléotide au hasard dans le génome de la levure (probabilités a priori définies ci-dessus). Quelle est la probabilité de tirer une purine (A ou G) ?
2. En supposant qu'une séquence est composée de résidus équiprobables, quelle est la probabilité du motif $GATWNA$ (W signifie "T ou A", et N correspond à n'importe quel nucléotide) ?
3. Même question en supposant les probabilités distinctes pour les nucléotides: $\mathrm{P}(A) = \mathrm{P}(T) = 0.3$, $\mathrm{P}(C) = \mathrm{P}(G)$ (à vous de calculer ces dernières).
4. En supposant des nucléotides équiprobables, quelle est la probabilités de n'observer aucun A dans un oligonucléotide de taille 12 ?