

Éléments d'analyse combinatoire

Probabilités et statistique pour la biologie (STAT1)

Jacques van Helden, revised by Lucie Khamvongsa-Charbonnier
and Yvon Mbouamboua

2018-11-16

Théorie

Résumé des concepts et formules

Résumé des concepts et formules

Théorie

Arrangements

On appelle **arrangements** les tirages *ordonnés* effectués *sans remise* au sein d'un ensemble.

Nombre d'arrangements de x éléments tirés parmi n .

$$A_n^x = \frac{n!}{(n-x)!} = \frac{n(n-1)\dots(n-x+1)(n-x)(n-x-1)\dots 2 \cdot 1}{(n-x)(n-x-1)\dots 2 \cdot 1} = n$$

Application typique: tiercé dans l'ordre. Les joueurs parient sur les trois chevaux gagnants d'une course ($x = 3$). Pour $n = 15$ chevaux partants, il existe $n \cdot (n-1) \cdot (n-2) = 15 \cdot 14 \cdot 13 = 2730$ possibilités.

Combinaisons

On appelle **combinaisons** le nombre de sous-ensembles qu'on peut tirer *sans remise* dans un ensemble, sans tenir compte de l'ordre des éléments tirés.

Ce nombre est fourni par le **coefficient binomial**.

$$\binom{n}{x} = C_n^x = \frac{n!}{x!(n-x)!}$$

Attention: les paramètres sont placés différemment dans la première (*binom* n x , “ x parmi n ”) et la seconde notation (C_n^x , “choose”).

Applications typiques:

- **tiercé** dans le désordre.

$$\binom{n}{x} \quad \binom{15}{3} \quad \frac{15!}{3!12!} = 455$$

Résumé des concepts et formules

Tirages avec / sans remise

Il existe deux types classiques de tirage d'éléments au sein d'un ensemble: avec ou sans remise.

1. **Tirage sans remise:** chaque élément peut être tiré au plus une fois. Exemples:
 - ▶ Jeu de loto (ou lotto).
 - ▶ Sélection aléatoire d'un ensemble de gènes dans un génome.
2. **Tirage avec remise:** chaque élément peut être tiré zéro, une ou plusieurs fois. Exemples:
 - ▶ Jeu de dés. A chaque lancer on dispose des mêmes possibilités (6 faces).
 - ▶ Génération d'une séquence aléatoire, par sélection itérative d'un élément dans l'ensemble des résidus (4 nucléotides pour l'ADN, 20 acides aminés pour les protéines).

Formules

Remise	Ordre	Formule	Description
Oui	Oui	n^x	Exponentielle : séquences de x éléments tirés dans un ensemble de taille n , avec remise.
Non	Oui	$n!$	Factorielle : toutes les permutations d'un ensemble de taille n
Non	Oui	$A_n^x = \frac{n!}{x!}$	Arrangements : listes (ordonnée) de x éléments tirés dans un ensemble de taille n
Non	Non	$C_n^x = \binom{n}{x} = \frac{n!}{x!(n-x)!}$	Combinaisons : ensembles (non ordonnés) de x éléments tirés dans un ensemble de taille n

Problème: dénombrement d'oligomères

L'ADN est composé de 4 nucléotides distincts dénotés par les lettres A, C, G, T, et les protéines de 20 acides aminés.

Pour chacun de ces deux types de polymères, combien d'oligomères distincts peut-on former en polymérisant 20 résidus ("20-mères") ?

Approche suggérée: simplifiez le problème au maximum, en commençant par des polymères beaucoup plus courts (1 résidu, 2 résidus).

Questions subsidiaires:

- ▶ Généralisez la formule pour les oligomères d'une longueur arbitraire k ("**k-mères**").
- ▶ Quel est le nom de la fonction donnant le résultat ?
- ▶ Dans ce processus, quel est le mode de sélection des résidus: **avec ou sans remise** ?

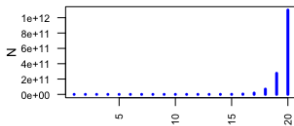
Solution: dénombrement d'oligomères

- ▶ Il s'agit d'un tirage avec remise: à chaque position de la séquence on a le choix entre n résidus (4 pour les acides nucléiques, 20 pour les protéines).
- ▶ Approche progressive de la solution
 - ▶ Cas trivial: séquence d'un seul résidu \rightarrow le nombre de possibilités est n .
 - ▶ Pour chacune des n possibilités de premier résidu, il y a n possibilités pour choisir le second résidu \rightarrow il existe $n \cdot n = n^2$ séquences de taille 2.
 - ▶ Pour chacune d'entre elles, n résidus possibles en 3^{ème} position \rightarrow il existe $n^2 \cdot n = n^3$ séquences distinctes de taille 3.
- ▶ Généralisation: il existe n^x séquences distinctes de taille x .
- ▶ Dans notre cas, la taille des séquences $x = 20$. On a donc
 - ▶ $N = n^x = 4^{20} = 1.1 \times 10^{12}$ oligonucléotides distincts
 - ▶ $N = n^x = 20^{20} = 1.05 \times 10^{26}$ oligopeptides distincts
- ▶ Il s'agit de la **distribution exponentielle**.

Nombre d'oligomères

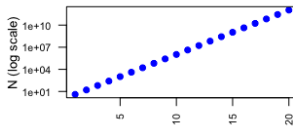
A FAIRE: quelques explications + les formules + graphique

Number of distinct oligonucleotides



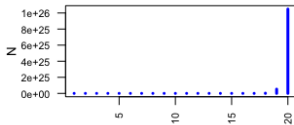
Oligomer size

Number of distinct oligonucleotides



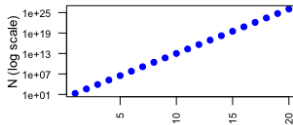
Oligomer size

Number of distinct oligopeptides



Oligomer size

Number of distinct oligopeptides



Oligomer size

Exercice: oligomères sans résidus répétés

Combien d'oligomères peut-on former (ADN ou peptides) en utilisant chaque résidu une et une seule fois ?

Approche suggérée: agrégez progressivement les résidus, en vous demandant à chaque étape combien d'entre eux n'ont pas encore été incorporés.

Questions subsidiaires:

- ▶ Généralisez la formule pour des séquences d'objets tirés dans un ensemble de taille arbitraire (n).
- ▶ Quel est le nom de la fonction donnant le résultat ?
- ▶ Dans ce processus, quel est le mode de sélection des résidus: **avec ou sans remise** ?

Solution: oligomères sans résidus répétés

- ▶ Premier résidu: n possibilités.
- ▶ Dès le moment où on a choisi ce premier résidu, il ne reste plus que $n - 1$ possibilités pour le second. On a donc $n \cdot (n - 1)$ possibilités pour les deux premiers résidus.
- ▶ Pour la troisième position, en fonction des deux premiers résidus déjà tirés, il ne reste que $n - 2$ résidus. On a donc $n \cdot (n - 1) \cdot (n - 2)$ possibilités pour les 3 premières positions de la séquence.
- ▶ Par extension, le nombre total de possibilités est donc (en supposant n suffisamment grand)

$$n! = n \cdot (n - 1) \cdot \dots \cdot 2 \cdot 1$$

- ▶ Dans notre cas:
 - ▶ $n! = 4! = 24$ oligonucléotides comportant exactement 1 fois chaque nucléotide (taille 4)

La factorielle

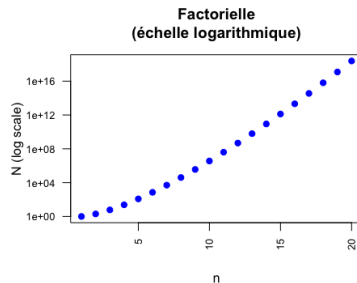
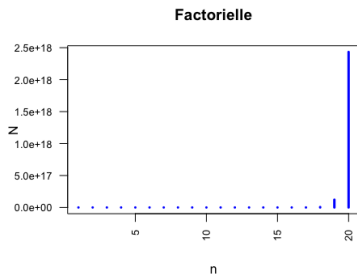
- ▶ S'applique pour dénombrer les permutations d'un ensemble.
- ▶ Il s'agit de tirages sans remise.
- ▶ Définie par une formule récursive.

$$N = n! = \begin{cases} 1 & \text{if } n = 0 \\ n \cdot (n-1)! & \text{otherwise} \end{cases}$$

Pour n suffisamment grand cela donne en clair.

$$N = n \cdot (n-1) \cdot (n-2) \dots 2 \cdot 1$$

Factorielle



Eléments de combinatoire

Nous résumons ici les formules les plus utilisées en analyse combinatoire.

- ▶ Arrangements (tirages ordonnés sans remise)
- ▶ Combinaisons (tirages non-ordonnés sans remise)

Arrangements

On appelle **arrangements** les tirages *ordonnés* effectués *sans remise* au sein d'un ensemble.

Nombre d'arrangements de x éléments tirés parmi n .

$$\begin{aligned}A_n^x &= \frac{n!}{(n-x)!} \\&= \frac{n(n-1)\dots(n-x+1)(n-x)(n-x-1)\dots 2 \cdot 1}{(n-x)(n-x-1)\dots 2 \cdot 1} \\&= n \cdot (n-1) \cdot \dots \cdot (n-x+1)\end{aligned}$$

Application typique:

- ▶ **tiercé** dans l'ordre.
- ▶ Les joueurs parient sur les trois chevaux gagnants d'une course ($x = 3$). Pour $n = 15$ chevaux partants, il existe $n \cdot (n-1) \cdot (n-2) = 15 \cdot 14 \cdot 13 = 2730$ possibilités.

Combinaisons

On appelle **combinaisons** le nombre de sous-ensembles qu'on peut tirer *sans remise* dans un ensemble, sans tenir compte de l'ordre des éléments tirés.

Ce nombre est fourni par le **coefficient binomial**.

$$\binom{n}{x} = C_n^x = \frac{n!}{x!(n-x)!}$$

Attention: les paramètres sont placés différemment dans la première (*binom* n x , “ x parmi n ”) et la seconde notation (C_n^x , “choose”).

Applications typiques:

- **tiercé** dans le désordre.

$$\binom{n}{x} \quad \binom{15}{3} \quad \frac{15!}{3!12!} = 455$$

Exercice: listes (ordonnées) de gènes

Lors d'une expérience de transcriptome indiquant le niveau d'expression de tous les gènes de la levure. Sachant que le génome comporte 6000 gènes, combien de possibilités existe-t-il pour sélectionner les 15 gènes les plus fortement exprimés (**en tenant compte** de l'ordre relatif de ces 15 gènes) ?

Approche suggérée: comme précédemment, simplifiez le problème en partant de la sélection minimale, et augmentez progressivement le nombre de gènes (1 gène, 2 gènes, ...).

Questions subsidiaires:

- ▶ Trouvez un exemple familier de jeu de pari apparenté à ce problème.
- ▶ Généralisez la formule pour la sélection d'une liste de x gènes dans un génome qui en comporte n .

Solution: listes (ordonnées) de gènes

Il s'agit d'une sélection **sans remise** (chaque gène apparaît à une et une seule position dans la liste de tous les gènes), et **ordonnée** (les mêmes gènes pris dans un ordre différent sont considérés comme un résultat différent).

- ▶ Pour le premier gène, il y a $n = 6000$ possibilités.
- ▶ Dès le moment où on connaît le premier gène, il n'existe plus que 5999 possibilités pour le second, et donc $n \cdot (n - 1) = 6000 \cdot 5999$ possibilités pour la suite des deux premiers gènes;
- ▶ Par extension, il existe $6000 \cdot 5999 \cdot 5998 \cdot \dots \cdot 5986 = 4.62 \times 10^{56}$ possibilités pour les 15 premiers gènes.
- ▶ En généralisant à la liste des x premiers gènes dans un ensemble de n , on obtient
$$N = n \cdot (n - 1) \cdot (n - 2) \cdot \dots \cdot (n - x + 1).$$

Exercice: ensembles (non-ordonnés) de gènes

Lors d'une expérience de transcriptome indiquant le niveau d'expression de tous les gènes de la levure. Sachant que le génome comporte 6000 gènes, combien de possibilité existe-t-il pour sélectionner les 15 gènes les plus fortement exprimés (**sans tenir compte** de l'ordre relatif de ces 15 gènes) ?

Approche suggérée: comme précédemment, simplifiez le problème en partant de sélections minimales (1 gène, 2 gènes, ...) et généralisez la formule.

Questions subsidiaires:

- ▶ Trouvez un exemple familier de jeu de pari apparenté à ce problème.
- ▶ Généralisez la formule pour la sélection d'un ensemble de x gènes dans un génome qui en comporte n .
- ▶ Connaissez-vous le nom de la formule ainsi trouvée ?

Solution: ensembles (non-ordonnés) de gènes

- ▶ Pour une sélection d'un seul gène, il existe $n = 6000$ possibilité.
- ▶ Pour 2 gènes, il existe $n \cdot (n - 1) = 6000 \cdot 5999$ arrangements, mais ceci inclut deux fois chaque paire de gènes $((a, b)$ et $(b, a))$. Le nombre d'ensembles non ordonnés est donc $N = n(n - 1)/2$.
- ▶ De même, pour 3 gènes, il faut diviser le nombre d'arrangements ($A_n^x = \frac{n!}{(n-x)!} = 6000 \cdot 5999 \cdot 5998$) par le nombre de permutations parmi tous les triplets de gènes $((a, b, c), (a, c, b), (b, a, c) \dots)$, ce qui donne $\frac{6000!}{(6000-3)!3!} = \frac{6000 \cdot 5999 \cdot 5888}{6} = 3.6 \times 10^{10}$.
- ▶ Pour 15 gènes, on obtient $\frac{n!}{(n-x)!x!} = \frac{6000!}{5985! \cdot 3!} = 3.53 \times 10^{44}$ combinaisons possibles.

Résumé des concepts et formules

Tirages avec / sans remise

Il existe deux types classiques de tirage d'éléments au sein d'un ensemble: avec ou sans remise.

1. **Tirage sans remise:** chaque élément peut être tiré au plus une fois. Exemples:
 - ▶ Jeu de loto (ou lotto).
 - ▶ Sélection aléatoire d'un ensemble de gènes dans un génome.
2. **Tirage avec remise:** chaque élément peut être tiré zéro, une ou plusieurs fois. Exemples:
 - ▶ Jeu de dés. A chaque lancer on dispose des mêmes possibilités (6 faces).
 - ▶ Génération d'une séquence aléatoire, par sélection itérative d'un élément dans l'ensemble des résidus (4 nucléotides pour l'ADN, 20 acides aminés pour les protéines).

Formules

Remise	Ordre	Formule	Description
Oui	Oui	n^x	Exponentielle : séquences de x éléments tirés dans un ensemble de taille n , avec remise.
Non	Oui	$n!$	Factorielle : toutes les permutations d'un ensemble de taille n
Non	Oui	$A_n^x = \frac{n!}{x!}$	Arrangements : listes (ordonnée) de x éléments tirés dans un ensemble de taille n
Non	Non	$C_n^x = \binom{n}{x} = \frac{n!}{x!(n-x)!}$	Combinaisons : ensembles (non ordonnés) de x éléments tirés dans un ensemble de taille n

Exercice : mutagénèse

On soumet un fragment d'ADN de 1 kilobase à un traitement mutagène qui provoque des mutations ponctuelles (substitutions) à 5 positions distinctes indépendantes. Combien de séquences possibles existe-t-il pour le fragment muté ?

Solution exercice 1: mutagenèse

On soumet un fragment d'ADN de 1 kilobase à un traitement mutagène qui provoque des mutations ponctuelles (substitutions) à 5 positions distinctes indépendantes. Combien de séquences possibles existe-t-il pour le fragment muté ?

Il s'agit de choisir au hasard 5 positions mutantes parmi les 1000 nucléotides du fragment d'ADN. Il s'agit d'un choix sans remise (chaque position ne peut être tirée qu'une fois), on choisit donc le coefficient binomial.

$$\binom{n}{x} = \binom{1000}{5} = C_{1000}^5 = \frac{1000!}{5!995!} = 8.2502913 \times 10^{12}$$

Exercice : oligopeptides 3×20

Combien d'oligopeptides de taille 60 peut-on former en utilisant exactement 3 fois chaque acide aminé ?

Solution de l'exercice : oligopeptides 3×20

Combien d'oligopeptides de taille 60 peut-on former en utilisant exactement 3 fois chaque acide aminé ?

Commençons par générer une séquence particulière qui remplit ces conditions, en concaténant 3 copies de chaque acide aminé, dans l'ordre alphabétique.

AAACCCDDDEEEFFFGGGHHHIIIKKKLLLMMMNNNPPPQQRRRSSSTTTVVVWWWYYY

Toutes les permutations de ces 60 lettres sont des solutions valides.
En voici trois exemples.

YMGALYHIRKTAKHAQFWGVWCFMCDIMVCLENFWVSKQHLSSDNIPEPPTNTEYQDRQ

FNGAAVEYICRQFHKTPWRNSSMTCHHMSKKGKPMGVPIEYLVIFQWDTEDRQYLDACLM

TKDMARVFAPCDCFYRHMITIEVTWGVKLNLSQYMLGEPRCHNWDPISWAFGENSHKQY