Solutions des exercices: probabilités et statistiques pour la biologie

Probabilités et statistique pour la biologie (STAT1)

Jacques van Helden

2018-01-05

Probabilité des longueurs d'ORF

Enoncé

On détecte les cadres ouverts de lecture (*open reading frames*, *ORF*) d'un génome en identifiant toutes les séquences de taille multiple de 3 comprises entre un start (ATG) et un stop (TAA, TAG ou TGA).

a. Sur base des fréquences génomiques de trinucléotides, calculer la probabilité de trouver à une position donnée du génome un ORF d'au moins 100 codons (plus précisément, la probabilité qu'à cette position précise commence un cadre ouvert de lecture d'au moins 100 codons).
b. Sachant que le génome fait 12 Mb, quel est le nombre attendu d'ORF d'au moins 100 codons ?

sequence	frequency	occurrences
AAA	0.039	478708
ATG	0.018	221902
TAA	0.022	272041
TAG	0.013	156668
TGA	0.020	244627

Approche

Cet exercice repose sur les concepts de probabilités combinées entre événements.

En particulier, nous mobiliserons les règles suivantes:

1. Des événements (A_1, A_2)sont mutuellement exclusifs si leur probabilité jointe est nulle ($P(A_1 \wedge A_2) = 0$).
2. La probabilité de l'union d'événements mutuellement exclusifs est la somme de leurs probabilités;

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

3. Des événements sont **complémentaires** s'ils sont mutuellement exclusifs et exhaustifs (ils couvrent l'ensemble des possibilités). La somme des probabilité d'événements complémentaires vaut 1.

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

4. 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.

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

Solution: probabilité des longueurs d'ORF

On sélectionne aléatoirement une position du génome (i). Nous allons raisonner en terme de codons (triplets de nucléotides): pour chaque position i on observe le triplet couvrant les nucléotides de i à $i+1$, et on avance par pas de 3 ($i, i+3, +6, \ldots$).

Pour y trouver le début d'un cadre ouvert de lecture d'au moins 100 codons,il faut remplir les conditions suivantes.

1. Présence d'un codon start (trinucléotide ATG) à la position i. La probabilité associée est directement fournie dans le tableau: $P(\mathrm{ATG}) = 0.018$.
2. Absence de codon stop (trinucléotides TAA, TAG ou TGA) pour les 99 triplets suivants (positions).

Probabilité d'absence du codon stop à une position donnée

L'absence d'un codon stop est l'événement *complémentaire* de la présence d'un condon stop (on observe exclusivement l'un ou l'autre). On en déduit que la somme des probabilités (absence ou présence) vaut 1.

$P(\neg\mathrm{STOP}) = 1 - P(\mathrm{STOP})$

Le symbole $\neg$ représente la négation logique (logical NOT).

Calculons la probabilité de **présence d'un codon stop** à une position donnée. Pour cela, il faut observer à cette position l'un des trois triplets suivants: TAG, TGA, TAA. Ces événements sont *mutuellement exclusifs*: à une position donnée, on en peut pas observer *à la fois* un TAA et un TGA. La probabilité d'observer l'un d'entre eux (probabilité de leur union) est donc la somme des probabilités individuelles.

$$\begin{aligned} P(\mathrm{STOP}) &= P(\mathrm{TAA}) + P(\mathrm{TAG}) + P(\mathrm{TGA}) \\ &= 0.022 + 0.013 + 0.02 = 0.055 \end{aligned}$$

On en détuit la probabilité d'**absence d'un codon stop** à une position donnée.

$$P(\neg\mathrm{STOP}) = 1 = P(\mathrm{STOP} = 1 - 0.06 = 0.94$$

Probabilité de présence d'ORF

Nous pouvons maintenant calculer la probabilité d'avoir un cadre ouvert de lecture d'au moins 100 codons qui commence à une position arbitraire i du génome.

Ceci correspond à la *probabilité jointe* (le ET logique) de l'ensemble des conditions requises.

- présence d'un codon start à la position i,
- ET absence de codon stop en position $i+3$,
- ET absence de codon stop en position $i+6$,
- ET ...
- ET absence de codon stop en position $i+197$.

On peut considérer que ces événements sont *indépendants* (le codon observé en position $i+3$ ne dépend pas de celui observé en position i). Leur probabilité jointe est donc le produit des probabilités individuelles.

$$
\begin{aligned}
P(\text{ORF}_{100}) &= P(\text{START}) \cdot \overbrace{P(\neg\text{STOP}) \cdot \ldots \cdot P(\neg\text{STOP})}^{99 \text{ times}} \\
&= P(\text{START}) \cdot P(\neg\text{STOP})^{99} \\
&= 0.018 \cdot 0.945^{99} = 6.653 \times 10^{-5}
\end{aligned}
$$

A priori cette probabilité n'a pas l'air énorme. Cependant, il faut tenir compte du fait qu'en annotant un génome on considère successivement toutes les positions possibles pour évaluer si on y trouve un cadre ouvert de lecture.

Nombre attendu d'ORF d'au moins 100 codons

Pour calculer le nombre d'ORF attendus au hasard, il faut multiplier la probabilité d'observer un ORF à une position donnée par le nombre de positions considérées.

Pour un génome de 12 Mb ($G = 1.2 \times 10^7$), on va considérer toutes les positions sur chacun des deux brins (on multiple donc par deux la taille du génome).

$$
E(\text{ORF}_{100}) = P(\text{ORF}_{100}) \cdot 2 \cdot G = 6.653 \times 10^{-5} \cdot 2 \cdot 1.2 \times 10^7 = 1596.7
$$

Interprétation du résultat

En 1996, lors de la première vague d'annotation du génome de la levure du boulanger, la stratégie d'annotation consistait à predire un gène codant chaque fois qu'on détectait un cadre ouvert de lecture d'au moins 100 codons. Les chercheurs étaient conscients du fait que cette stratégie était susceptible de produire un nombre important de fausses prédictions, mais il s'agissait du tout premier génome eucaryote séquencé, et il fallait bien commencr par quelque chose pour se donner une chance d'y découvrir de nouveaux gènes.

Nous venons de calculer le nombre attendu de faux-positifs dans ce processus: si on avait généré une séquence aléatoire de la même taille selon un modèle de Markov respectant les mêmes fréquences de trinucléotides, on s'attendrait à y trouver 1596.7 cadres ouverts de lecture d'au moins 100 codons. Ceci indique que les gènes de taille relativement courte (300 nucléotides) devaient être considérés avec circonspection, et soumis à analyse ultérieure avant d'être annotés comme des gènes fiables.

En 2003, une stratégie complémentaire a été mise à contribution pour évaluer la fiabilité des ORF prédits dans le génome de *Saccharomyces cerevisiae*, en s'appuyant sur la génomique comparative: après avoir séquençé les génomes de quelques autres espèces du même genre (*Saccharomyces*), on a testé la conservation des ORF initialement prédits, en supposant que si on trouvait des ORF similaires dans les 4 espèces, cela suggérait une pression sélective positive pour l'absence de codon stop dans ces régions, et renforçait la prédiction d'un gène codant correspondant à ce cadre ouvert de lecture.

Probabilité d'un motif avec erreurs

On recherche dans un génome les occurrences du motif GATAAG en admettant un certain nombre de substitutions. En supposant que les nucléotides sont indépendants et équiprobables, quelle est la probabilité de trouver à une position du génome.

a. Une instance exacte du motif (aucune substitution) ?
b. Une séquence ne présentant aucune correspondance avec le motif (6 substitutions) ?
c. Une instance avec exactement 1 substitution ?
d. Une instance avec au plus 2 substitutions ?

Cadre théorique

On modélise le problème comme un **schéma de Bernoulli**: chaque position de la séquence correspond à un essai, qui peut donner soit un *succès* (correspondance avec motif) soit un *échec* (différence avec le motif), et on suppose que la probabilité de succès est constante (nucléotides indépendants et équiprobables).

Le motif GATAAG fait 6 nucléotides, on aura donc 6 essais successifs, qui viseront à tester successivement l'identité entre le premier, second, ... sixième nucléotide de la séquence et le nucléotide à la position correspondante du motif. Pour chaque position, on a une probabilité de succès $p = 1/4$ et d'échec $q = 1 - p = 3/4$.

Probabilité d'nstance exacte du motif (aucune substitution)

Pour avoir une instance exacte du motif, il faut une succession de 6 correspondances entre nucléotides.

$$P(\text{perfect match}) = p^6 = 0.25^6 = 2.4 \times 10^{-4}$$

Probabilité d'une séquence ne présentant aucune correspondance avec le motif (6 substitutions)

$$P(\text{full mismatch}) = (1-p)^6 = 0.75^6 = 0.178$$

Probabilité d'une instance avec exactement 1 substitution

On peut trouver la réponse par un raisonnement assez simple. Les instances avec 1 substitutions combinent - 5 "succès"" (correspondances) ayant une probabilité $p = 0.25$, - 1 "échec" (substitition) $q = 1 - p = 0.75$.

La substitution peut se trouver à n'importe quelle position du motif. Si l'on représente respectivement par 1 et 0 les succès et échecs, on acceptera les disposition suivantes de succès et d'échecs.

Disposition	Probabilité
111110	$p \cdot p \cdot p \cdot p \cdot p \cdot (1-p) = p^5(1-p)$
111101	$p \cdot p \cdot p \cdot p \cdot (1-p) \cdot p = p^5(1-p)$
111011	$p \cdot p \cdot p \cdot (1-p) \cdot p \cdot p = p^5(1-p)$
110111	$p \cdot p \cdot (1-p) \cdot p \cdot p \cdot p = p^5(1-p)$
101111	$p \cdot (1-p) \cdot p \cdot p \cdot p \cdot p = p^5(1-p)$
011111	$(1-p) \cdot p \cdot p \cdot p \cdot p \cdot p = p^5(1-p)$

Ces dispositions sont mutuellement exclusive (si on a exactement une substitution, elle peut se trouver soit à la première soit à la deuxième position, mais pas aux deux en même temps). On peut donc calculer la prbabilité de l'union des 6 dispositions comme la somme de leurs probabilités.

$$P(\text{match withb 1 subst}) = 6 \cdot p^5 \cdot (1-p) = 6 \cdot 0.25^5 \cdot 0.75 = 0.0044$$

Un raisonnement plus rapide: il s'agit d'un schéma de Bernoulli avec 6 essais indépendant, ayant la même probabilité de succès $p = 0.25$. La probabilité d'observer exactement $x = 5$ succès parmi $n = 6$ essais est donnée par la **distribution binomiale**.

$$
\begin{aligned}
P(x = 5|n = 6, p = 0.25) &= \binom{n}{x} p^x (1-p)^{n-x} \\
&= \frac{n}{x!(n-x)!} p^x (1-p)^{n-x} \\
&= \frac{6}{5! \cdot 1!} 0.25^5 \cdot 0.75 = 0.0044
\end{aligned}
$$

Probabilité d'une instance avec au plus 2 substitutions

Pour avoir au plus deux substitutions, il faut avoir soit 0, soir 1, soit 2 subsitution. Ces possibilités sont mutuellement exclusives (on ne peut avoir à la fois 1 et 2 substitutions), la probabilité de leur union est donc la somme de leurs probabilités.

Nous avons calculé ci-dessus la probabilité d'avoir 0 et 1 substitution. Par le même raisonnement, on peut calculer la prbabilité de 2 substitutions avec le distribution binomiale.

$$
\begin{aligned}
P(x = 4|n = 6, p = 0.25) &= \binom{n}{x} p^x (1-p)^{n-x} \\
&= \frac{n}{x!(n-x)!} p^x (1-p)^{n-x} \\
&= \frac{6}{4! \cdot 2!} 0.25^4 \cdot 0.75^2 = 0.033
\end{aligned}
$$

La solution peut s'écrire sous la forme de 3 probabilités binomiales correspondant respectivement à 0, 1 et 2 substitutions, ou, de façon équivalente, à 4, 5 ou 6 succès ($x \geq 4$).

Note: afin de vous familiariser avec les deux formulations alternatives du coefficient binomial, nous utilisons ici la formule "choose" C_n^x, qui est équivalente à $\binom{n}{x}$ (notez l'inversion des positions de x et n).

$$
\begin{aligned}
P(i \geq 4|n = 6, p = 0.25) &= \sum_{i=4}^{n} C_n^i p^i (1-p)^{n-i} \\
&= \sum_{i=4}^{6} \frac{6}{i! \cdot (n-i)!} 0.25^i \cdot 0.75^{n-i} \\
&= 0.033 + 0.0044 + 2.4 \times 10^{-4} \\
&= 0.038
\end{aligned}
$$

Alignement de lectures NGS sans erreur

Au terme d'un séquençage de type "Next Generation Sequencing" (NGS), on dispose d'une librairie de $N = 10^6$ lectures courtes. On aligne la librairie sur le génome de référence, dont la somme des chromosomes fait 1 gigabase ($G = 10^9$ paires de bases), en utilisant un algorithme d'alignement sans gap.

Calculez la distribution de probabilité d'obtenir un alignement sans erreur en fonction de la longueur des lectures (k).

Cadre théorique

Schéma de Bernoulli: on suppose que les nucléotides se succèdent de façon indépendante et ont une probabilité constante $p = 1/4$ (équiprobabilité).

Probabilité d'alignement sans erreur

La probabilité de n'obtenir aucune erreur sur k nucléotides vaut.

$P(X = k) = p^k$

k	Proba.match
1	2.50e-01
2	6.25e-02
3	1.56e-02
4	3.91e-03
5	9.77e-04
6	2.44e-04
7	6.10e-05
8	1.53e-05
9	3.81e-06
10	9.54e-07

Alignement de lecture NGS avec erreurs

Un biologiste a fait séquencer un échantillon et a obtenu un fichier comportant 50 millions de lectures (« short reads ») de 35 paires de base, qu'il aligne sur le génome humain (3 gigabases répartis sur 23 chromosomes). Durant l'alignement, il choisit d'accepter au maximum 3 substitutions par lecture.

a. En supposant un modèle de fond basé sur des nucléotides équiprobables et distribués de façon indépendante, comment calculeriez-vous la probabilité pour qu'un read s'aligne complètement à une position arbitraire du génome, avec au plus 3 substitutions (sans indel). Indiquez la formule et justifiez votre choix.

b. Sous ces mêmes conditions, quel serait le nombre de faux-positifs attendus si l'on aligne l'ensemble de la librairie de séquences sur l'ensemble du génome ?

a. Probabilité d'alignement avec au plus 3 substitutions

On modélise l'alignement(sans gap) d'un read à une position donnée du génome comme un schéma de Bernoulli, où chaque position alignée (nucléotide) correspond à un essai, avec une probabilité de succès $p = 0.25$ (identité entre le nucléotide du read et celui du génome) et d'échec $q = 1 - p = 0.75$ (différence entre les deux nucléotides alignés), et où le nombre d'essais correspond à la longueur des reads ($k = 35$).

La probabilité d'observer au plus 3 substitutions est la somme des probabilités d'observer exactement 0, 1, 2 ou 3 substitutions.

Chacune de ces probabilités peut s'obtenir avec la loi de densité binomiale.

$$P(X = i) = C_k^i p^i (1 - p)^{k-i}$$

m étant le nombre de différences (mismatches), le nombre d'identités vaut donc $i = k - m$. i prendra donc successivement les valeurs de 32 à 35.

La probabilité de l'alignement avec au plus 3 substitutions vaut donc.

$$P = P(M \leq 3) = P(X \geq k-3) = \sum_{i=k-3}^{k} C_k^i \cdot p^i \cdot (1-p)^{k-i}$$
$$= \sum_{i=32}^{35} C_{35}^i \cdot 0.25^i \cdot 0.75^{35-i}$$
$$= 1.54 \times 10^{-16}$$

b. Nombre de faux-positifs attendus en alignant l'ensemble de la librairie sur le génome complet

Le nombre de faux positifs attendus (E) est la p-valeur (P) calculée ci-dessus multiplée par le nombre de tests. En effet, on tente d'aligner chacun des 5×10^7 fragments de lecture ("read") sur les 2 brins de chacune des 3×10^9 positions du génome.

Cependant, il faut tenir compte du fait qu'un read de longueur $k = 35$ ne peut pas être aligné sur les $k-1$ positions d'un chromosome, puisque l'alignement serait incomplet (rappelons qu'on effectue ici des alignements globaux, qui doivent donc couvrir l'ensemble de la longueur du read). Il faut donc soustraire $k - 1 = 34$ positions sur chaque brin de chaque chromosome, donc au total $2 \cdot C \cdot (k-1)$ positions.

Sites de restriction

Dans un génome bactérien de 4 Mb avec une composition de 50% de G+C, on observe 130 occurrences de l'hexanucléotide GGCGCC. On suppose un schéma de Bernoulli et une composition équiprobable de nucléotides.

a. Quelle est la probabilité d'observer une occurrence de GGCGCC à une position donnée du génome ?

b. Combien d'occurrences s'attend-on à trouver dans l'ensemble du génome ?

c. Quelle serait la probabilité d'observer un nombre aussi faible d'occurrences (130 ou moins) si l'on générait une séquence aléatoire selon le modèle de Bernoulli avec nucléotides équiprobables ?

d. Comment peut-on interpréter cette sous-représentation de l'hexanucléotide GGCGCC du point de vue biologique ?

Probabilité d'occurrence

$$P(\text{GGCGCC}) = p^k = 0.25^6 = 2.44 \times 10^{-4}$$

Occurrences attendues sur l'ensemble du génome

On multiplie la probabilité d'occurrence à une position donnée par le nombre de positions génomiques.

Comme le site de restriction est réverse complémentaire palindromique, chaque occurrence sur un brin est nécessairement associée à une occurrence sur l'autre brin. Pour connaître le nombre de positions génomiques susceptibles d'être reconnues par l'enzyme de restriction, il suffit donc de compter les occurrences sur un brin du génome.

$$E(\text{GGCGCC}) = P(\text{GGCGCC}) \cdot G = 2.44 \times 10^{-4} \cdot G = 976.6$$

Probabilité d'observer au plus 130 occurrences

Le nombre d'occurrences observées ($x = 130$) semble nettement inférieur à l'espérance ($E(\text{GGCGCC}) = 976.6$), c'est-à-dire le nombre attendu étant donné le modèle de background. Nous allons maintenant évaluer la significativité statistique de cette observation en calculant la probabilité d'observer une telle différence par hasard. Si cette probabilité est très faible, on considérera le résultat comme significatif et on tentera d'en fournir une interprétation biologique.

Selon le modèle null (séquences générées au hasard avec des nucléotides équiprobables) on peut modéliser les occurrences d'un site de restriction comme un schéma de Bernoulli: à chaque position génomique, on observe un site (*succès*) ou non (*échec*), avec une probabilité constante $P(\text{GGCGCC}) = 2.44 \times 10^{-4}$.

Par simplification, on considérera que la présence/absence d'un site à une position i est indépendante de sa présence/absence à la position $i - 1$ (l'undépendance entre essais successifs est une des conditions du modèle de Bernoulli).

Notons qu'il s'agit là d'une simplification abusive, car si l'on observe un site GGCGCC à une position donnée, il est impossible d'en observer une autre occurrence sur les 5 positions suivantes. Puisque le génome comporte $x = 130$ occurrences du site de restriction, on sait donc qu'il ext exclus de trouver un site sur les $5 \cdot 130 = 650$ positions qui chevauchent ces instances. Il existe des méthodes permettant de modéliser de telles dépendances, mais elles reposent sur des modélisations trop avancées pour ce cours. Dans l'immédiat, nous adopterons une solution pratique, en considérant que ces 650 positions exclues ne représentent qu'une fraction négligeable du génome ($G = 4 \times 10^6$).

Après avoir jusitifié (toute mesure gardée) notre modèle de Bernoulli, nous pouvons calculer la probabilité du nombre de sites de restriction avec la **loi binomiale**. Nous calculons ici la surface de la **queue à gauche** de l'observation: $P(X \leq 130)$, c'est-à-dire la probabilité d'observer entre 0 et 130 occurrences.

$$\begin{aligned} P(X \leq x) &= \sum_{i=0}^{x} C_G^i P(\text{GGCGCC})^i \cdot (1 - P(\text{GGCGCC}))^{G-i} \\ &= \sum_{i=0}^{130} C_{4\times10^6}^i \cdot (2.44 \times 10^{-4})^i \cdot \left(1 - 2.44 \times 10^{-4}\right)^{4\times10^6 - i} \\ &= 7.61 \times 10^{-257} \end{aligned}$$

Cette probabilité est extrêmement faible, il serait donc extrêmement improbable d'obtenir un aussi petit nombre d'instances du motif GGCGCC dans un génome de même taille généré aléatoirement.

Interprétation biologique

Le très faible nombre d'occurrences du site GGCGCC se comprend parfaitement dès qu'on sait que ce site est reconnu par une enzyme de restriction, dont l'activité moléculaire est de cliver l'ADN. Ces enzymes jouent un rôle de protection en fragmentant les molécules d'ADN étrangères au génome de *Escherichia coli*, mais elle pourrait également endommager le génome propre de la bactérie. Le très faible nombre de sites pour sa propre enzyme de restriction résulte d'un effet de contre-sélection de ces sites dans le génome d'*Escherichia coli*.

Détection de gènes différentiellement exprimés

Un chercheur a analysé, à l'aide de biopuces, le niveau d'expression de l'ensemble des gènes à partir d'échantillons sanguins prélevés chez 50 patients ($n_p = 50$) et chez 50 sujets témoins ($n_t = 50$). Il s'intéresse

particulièrement à un gène qui semble montrer une différence entre les 2 groupes. Ainsi, il ré-analyse l'expression du même gène dans les mêmes échantillons en utilisant une autre technique, la qPCR. Il obtient

- pour les patients, une moyenne $m_p = 21$
- pour les contrôles, une moyenne $m_t = 10$
- des écarts-types identiques pour les 2 groupes $s_p = s_t = s = 15$

Afin de tester si la différence observée entre les moyennes est significative, le chercheur décide d'effectuer un test de Student.

a. Le choix du test de Student vous semble-t-il approprié ? Justifiez le choix du chercheur. Quelles auraient été les alternatives envisageables ?

b. Sachant qu'a priori on ne sait pas dans quel sens la maladie pourrait affecter le niveau d'expression du gène, préconisez-vous un test uni- ou bilatéral ?

c. Formulez l'hypothèse nulle et expliquez-la en une phrase.

d. Sur base de la formule ci-dessous, calculez la statistique t de Student.

e. Indiquez, en vous basant sur les tables fournies, la p-valeur correspondante.

f. Interprétez la p-valeur, et aidez le chercheur à tirer les conclusions de son étude.

Choix du test

Le test de Student repose sur deux hypothèses de travail: normalité des données et homoscédasticité.

La première hypothèse de travail du test de Student est la normalité dans chacune des deux populations dont l'échantillon est extrait. A priori on ne peut pas garantir que les données d'expression suivent une distribution normale. Le choix d'un test paramétrique pourrait donc être mis en question. On pourrait envisager d'effectuer, préalablement au test de comparaison de moyenne, un test de normalité (séparément pour chaque gène et dans chaque groupe). Si le test s'avère positif (rejet de l'hypothèse de normalité), on optera pour un test non-paramétrique (Mann-Whitney-Wilcoxon).

Toutefois, le test de normalité ne serait sans doute pas très informatif, car avec un effectif de 50 par test on aurait une trop faible puissance pour pouvoir détecter de écarts à la normalité. Cependant, nous savons que les tests paramétriques de Student et de Welch sont relativement robustes à la non-normalité quand la taille de l'échantillon est suffisamment grande (typiquement $n > 30$).

La seconde hypothèse de travail est que les populations d'où proviennent les échantillons ont la même variance (homoscédasticité). En cas d'homoscédasticité on peut appliquer le test de Student, et dans le cas contraire on recourt au test de Welch. Comme nous observons des écarts-types identiques pour les deux échantillons, nous pouvons considérer que cette hypothèse de travail est valide.

Dans le cas présent, nous pouvons donc appliquer un test de Student.

Orientation du test

On appliquera un test bilatéral, afin de détecter un impact sur l'expression, qu'il s'agisse d'une augmentation ou d'une diminution.

Formulez l'hypothèse nulle et expliquez-la en une phrase

$$H_0 : \mu_p = \mu_t$$

Nous testons l'égalité des moyennes d'expression de ce gène entre les populations de patients (μ_p) et de témoins (μ_t).

Il faut noter qu'on représente ici les moyennes de populations par le symbole grec μ ("mu"), alors que les moyennes d'échantillons sont représentées par la lettre latine m. Les hypothèses à tester portent toujours sur les paramètres de la population, et non sur ceux des échantillons (les moyennes d'échantillons sont différentes, puisque $21 \neq 10$).

Calcul de la statistique t de Student

$$t = \frac{m_p - m_t}{s \cdot \sqrt{\frac{1}{n_p} + \frac{1}{n_t}}} = \frac{21 - 10}{15 \cdot \sqrt{\frac{1}{50} + \frac{1}{50}}} = \frac{11}{\frac{15}{5}} = \frac{11}{3} = 3.667$$

Extraction de la p-valeur à partir de la table de Student

Nous nous basons sur la table de t, avec $\nu = n_1 + n_2 \breve{} 2 = 98$ degrés de liberté (dans la table fournie, nous utiliserons la ligne correspondant à 100 degrés de libertés). La valeur t la plus élevée dans la table est $t = 3.390$, qui pour le test bilatéral correspond à une p-valeur de 0.001. Nous pouvons donc conclure que la p-valeur correspondant à une statistique t de 3.667 est inférieure à 0.001.

Note: avec **R** nous pouvons calculer précisément la p-valeur associée à cette valeur de la statistique de Student: $P(t \geq 3.667) = 3.86 \times 10^{-4}$.

Interprétation de la p-valeur et décision

La P-valeur d'un test d'hypothèse indique la probabilité d'obtenir un résultat au moins aussi extrême que celui observé, si l'on était sous hypothèse nulle (autrement dit si les échantillons avaient été tirés de populations ayant la même moyenne). Comme il s'agit d'un test bilatéral, on considérera comme extrêmes les deux queues de la distribution de Student.

$$P = P((T \geq 3.667) \vee (T \geq -3.667)) = 2 \cdot P(T \geq 3.667)$$

Une P-valeur très faible indique qu'il est très peu probable que des échantillons tirés de deux populations de même moyenne diffèrent autant.

Comme le test ne s'applique qu'à un seul gène (celui pour lequel on aeffectué l'expérience de PCR), une valeur $P < 0.001$ peut être considérée comme très significative, et l'on peut rejeter l'hypothèse nulle. On peut donc considérer que ce gène est différentiellement exprimé entre les deux populations.

Taux d'anticorps

Un groupe de chercheurs a détecté l'association, avec la résistance à la bilharziose, de taux élevés d'IgE spécifiques, une classe particulière d'anticorps. D'autres chercheurs ont cherché à répliquer ce résultat dans une population indépendante exposée à la bilharziose. Les résultats obtenus sont indiqués ci-dessous.

- Pour les sujets résistants ($n_r = 32$), la moyenne $m_r = 10$.
- Pour les sujets susceptibles ($n_s = 32$), la moyenne $m_s = 7$.
- Les écarts-types des deux groupes sont égaux : $s_r = s_s = s = 2.8$.

a. Quelle méthode préconisez-vous pour tester l'égalité des moyennes (justifiez) ? Quelles sont les hypothèses de travail de ce test ?

b. En partant du principe que ces conditions sont remplies dans le cas présent, formulez l'hypothèse nulle et calculez le score t de Student (formule ci-dessous). Enfin, estimez P valeur à partir de la table fournie.

c. A l'issue du test, quelle décision prenez-vous ? Justifiez votre réponse.

Méthode préconisée pour tester l'égalité des moyennes

Pour choisir un test d'égalité des moyennes, la première question est de savoir si l'on peut ou non appliquer un test paramétrique.

A priori nnous ne disposons pas d'information concernant la distribution des taux d'IgE dans les populations résistantes et susceptibles. Nous ne pouvons donc pas présupposer que les données suivent une distribution normale.

Cependant, les deux échantillons sont de taille suffisante ($n_1 = n_2 = 32$) pour qu'on puisse s'affranchir de cette condition. En effet, même si les données ne suivent pas une distribution normale, la moyenne d'échantillons tend vers la normalité quand l'effectif augmente, et, de façon pragmatique, on considère généralement qu'au-delà de 30 éléments par échantillon on peut appliquer un test paramétrique.

Parmi les tests paramétriques, on choisira soit un test t de Student si les deux populations sont de même variance (*homoscédasticité*), soit un test t de Welch si leurs variances diffèrent (*hétéroscédasticité*). Comme dans notre cas les écarts-types des deux échantillons sont identiques, on peut présupposer que les populations dont ils sont extraits ont la même variance (en tout état de cause, si l'on réalisait un test d'égalité des variances, on serait amené à accepter l'hypothèse nulle).

Hypothèse nulle

On effectue un **test unilatéral** puisqu'on teste specifiquement l'association de la résistance à un taux élevé d'anticorps.

$$H_0 : \mu_R \leq \mu_S H_1 : \mu_R > \mu_S$$

où μ_R et μ_S sont respectivement les taux moyens d'anticorps chez les individus résistants et sensibles.

Calcul du score t de Student

$$\begin{aligned} t_S = \frac{\hat{\delta}}{\hat{\sigma}_\delta} &= \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \\ &= \frac{10 - 7}{\sqrt{\frac{32 \cdot 2.8^2 + 32 \cdot 2.8^2}{32 + 32 - 2}\left(\frac{1}{32} + \frac{1}{32}\right)}} \\ &= \frac{10 - 7}{2.8 \cdot \sqrt{\frac{64}{62}}\sqrt{\frac{1}{32} + \frac{1}{32}}} \\ &= \frac{3}{2.8 \cdot 1.016 \cdot \frac{1}{4}} = 4.22 \end{aligned}$$

Estimation de la probabilité critique (p valeur)

On se base sur le tableau de la distribution t de Student (disponible pour l'examen), en considérant le test unilatéral (*one-tail*).

On sélectionne la ligne correspondant aux degrés de liberté $\mathrm{df} = n_1 + n_2 - 2 = 62$). Comme le tableau ne mentionne pas cette valeur, on choisit la ligne la plus proche ($\mathrm{df} = 60$).

La valeur t observée ($t_{obs} = 4.22$) est supérieure à la plus grande valeur indiquée dans la table ($t = 3.460$, pour $p = 0.0005$). On peut donc conclure que la probabilité critique est inférieure à 0.0005.

Décision et jusification

Pour un test isolé, une probabilité critique de $p = 0.0005$ signifie que sous hypothèse nulle – c'est-à-dire si la résistance n'était pas associée à un taux élevé d'anticorps – la probabilité d'obsrver une différence de moyennes aussi grande serait de 0.0005. Cette probabilité est très faible, on peut donc rejeter l'hypothèse nulle, et conclure que la différence observée est significative.

Enrichissement fonctionnel

Dans le génome de la levure, 40 gènes ont été assignés à la classe fonctionnelle "Biological Process: Methionin Biosynthesis". Une expérience de transcriptome rapporte 80 gènes différentiellement exprimés, dont 10 appartiennent à cette classe fonctionnelle. Sachant que le génome comporte 6000 gènes, peut-on consiérer ce résultat comme significatif ?

Pas vu en 2016-2017.

Jeu de roulette

La roulette comporte 37 nombres allant du 0 au 36. Un joueur a décidé de miser systématiquement 1 euro sur le nombre 17 jusqu'à ce que ce nombre sorte, et de s'arrêter ensuite. Sachant que quand on mise sur un seul nombre, le gain vaut 36 fois la mise, quelle est la probabilité pour que le joueur sorte du casino en ayant gagné de l'argent ? Il n'est pas nécessaire de fournir une réponse numérique, vous pouvez vous contenter d'indiquer la formule, en indiquant les nombres correspondant aux différents symboles mathématiques. Justifiez votre réponse en expliquant votre raisonnement.

Cadre théorique

Il s'agit de calculer le temps d'attente du premier succès dans un schéma de Bernoulli. Chaque pari que le joueur place sur le nombre 17 peut être considéré comme un essai qui peut résulter soit en un succès (le nombre 17 sort) soit en un échec (un autre nombre sort).

La (petite) difficulté est de coupler la réflexion sur ce temps d'attente avec un calcul des coûts et bénéfices.

Temps d'attente du premier succès

A chaque essai, la probabilité de succès vaut $p = 1/37 = 0.027$. La probabilité d'obtenir le premier succès au $x^{\text{ième}}$ essai se calcule selon la **loi géométrique**.

$$P(X = x) = (1 - p)^{x-1}p$$

En effet pour que le premier succès se produise au $x^{\text{ième}}$ essai, il faut commencer par $x - 1$ essais infructuex (avec une probabilité $q = 1 - p$) et terminer par un essai fructueux (de probabilité p).

Fonction cumulative de distribution

La fonction cumulative de distribution (***CDF***) indique la probabilité pour que le pemier succès ait lieu au $x^{\text{ième}}$ essai ou avant. Pour la loi géométrique, elle s'obtient aisément en raisonnant sur le complément: la probabilité de n'obtenir aucun succès durant les x premiers essais vaut $(1-p)^x$. La probabilité pour que le premier succès sorte au plus tard au $x^{\text{ième}}$ essai vaut donc.

$$P(X \leq x) = 1 - (1-p)^x$$

Calcul des coûts / bénéfices

A chaque essai, le joueur mise un euro. Durant les $x-1$ premiers essais (ceux où il échoue), ses pertes cumulées s'élèvent donc à $(x-1) \cdot 1€$. Au $x^{\text{ième}}$ essai, il remporte 36€ et arrête le jeu.

Si le numéro 17 sort avant le 36ème essai, le joueur sortira gagnant. S'il sort au 36ème essai, il ne sera ni gagnant ni perdant, et après le 36ème essai il sera perdant.

On peut facilement calculer la probabilité de ne pas sortir gagnant c'est-à-dire de n'obtenir aucun succès durant les 35 premiers essais.

$$P(x \geq 36) = (1-p)^{35} = 0.383$$

La probabilité d'obtenir un premier succès avant le 36ème essai est le complément de cette probabilité :

$$P(X < 36) = 1^{\smile} P(x \geq 36) = 1^{\smile} (1-p)^{35} = 1^{\smile} \left(\frac{36}{37}\right)^{35} = 0.617$$

Information complémentaire : le fait que la probabilité de sortir gagnant soit supérieure à 50% indique que le joueur a plus de chances de gagner que de perdre, ce qui semble paradoxal puisque si c'était le cas le casino jouerait perdant. Cependant, il faut également prendre en compte le montant des gains et des pertes en fonction du temps d'attente (ceci ne faisait pas partie de la question).

Certes, s'ils suivent la procédure indiquée (miser systématiquement 1€ jusqu'au premier succès), la majorité des joueurs sortent gagnants (61.7% voient leur nombre sortir avant la 36ème tentative), mais ***le gain moyen est négatif***, car les sommes perçues auprès des perdants (les 38.3% de joueurs dont le nombre ne sort pas avant le $36^{\text{ème}}$ essai) dépassent les bénéfices des gagnants.

On peut démontrer que le bénéfice attendu est de -1 (perte de 1€). Autrement dit, si un grand nombre de joueurs adoptent cette stratégie (miser 1 euro sur un nombre particulier et s'arrêter dès le premier succès) le casino gagne en moyenne un euro par joueur. Comme pour tous les jeux de hasard, le gagnant systématique est le casino.

Concepts de probabilité

En quoi consiste le modèle de Bernoulli ? Ce modèle est-il généralement adapté à l'analyse des séquences biologiques ? Justifiez en quelques phrases.

Définiton du modèle de Bernoulli

La définition générale du schéma de Bernoulli repose sur une série d'essais (expériences) indpendants, pouvant chacun déboucher sur deux résultats mutuellement exclusifs, respectivement qualifiés de "succès" et "échec", avec une probabilité de succès (p) identique pour chaque essai.

En analyse de séquences, on se base souvent sur un extension du modèle, appelée modèle de Bernoulli généralisé, où chaque essai résulte en un résultat parmi un nombre fini de possibilités. Pour le séquences nucléiques on considère 4 possibilités correspondant aux nucléotides, et pour les séquences peptidiques 20 possibilités correspondant aux acides aminés. Le modèle de Bernoulli généralisé postule que les résidus (nucléotides ou acidés aminés selon le type de séquence) se succèdent de façon indépendante, et que la probabilité d'un résidu donné est la même à chaque position de la séquence.

Attention, ceci ne signifie pas forcément que tous les résidus ont la même probabilité. On peut définir un modèle de Bernoulli où chaque résidu a une probabilité spécifique (par exemple une séquence d'ADN avec $P_A = 0.31, P_T = 0.29, P_C = 0.18, P_G = 0.22$). Si ces probabilités sont considérées comme identiques à chaque position de la séquence, on pourra parler de modèle de Bernoulli.

Adéquation à l'analyse des séquences biologiques

Les modèles de Beroulli présentent des qualités pratiques mais il est important de connaître leurs limitations. En effet, l'hypothèse d'indépendance entre résidus successifs n'est pas réaliste pour des séquences macromoléculaires. A titre d'exemples:

- Dans les génomes de mammifères, le dinucléotde CG (également dénoté CpG, où p indique le groupe phosphate) est beaucoup moins fréquent que ce à quoi on s'attendrait sous hypothèse d'indépendance: $F_{CG} < F_C \cdot F_G$. Ceci résulte du fait que ce dinucléotide est fréquemment méthylé par des enzymes spécifiques, qui jouent un rôle dans l'inactivation de l'ADN et la reconnaissance de l'ADN viral. La méthylation de CpG favorise la mutation vers une thymine, par déamination, ce qui a pu susciter une réduction progressive de leur fréquence au cours de l'évolution des mammifères.
- Les séquences non-codantes sont enrichis en oligonucléotides riches en adénines et thymines (poly-AT) par rapport à la fréquence à laquelle on s'attendrait d'après un modèle de Bernoulli.
- Les successions d'acides aminés sont en partie influencées par leurs propriétés biochimiques (par exemple, on trouvera souvent des successions d'acides aminés hydrophobes dans les domaines transmembranaires).
