PROBABILITES ET STATISTIQUES POUR LA BIOLOGIE
(STAT1, ENSBBAU16L) – EXAMEN – 22 JANVIER 2018

Calculatrices Autorisées ; Documents Non Autorisés.

Pondération : cet examen compte pour 100% de la note finale.

Question 1 (4 points)

On fait une recherche de similarités avec l'algorithme BLAST, en comparant une séquence protéique inconnue aux 25000 séquences d'un protéome de référence. On décide arbitrairement de retenir les 5 meilleures correspondances, et d'interpréter le résultat en tenant compte de leur ordre. Combien de possibilités y a-t-il ? Expliquez le raisonnement, indiquez le nom de la fonction et sa formule (avec les symboles), puis remplacez dans cette formule chacun des symboles par les valeurs numériques appropriées de l'énoncé. Il n'est pas nécessaire de fournir le résultat numérique final, dont le calcul nécessiterait un ordinateur.

Il s'agit d'un tirage ordonné sans remise, il s'agit donc d'un arrangement (choix de $x=5$ gènes parmi $n=25000$).

$$A_n^x = \frac{n!}{(n-x)!} = A_{25.000}^{27} = \frac{25.000!}{24.973!} = 25.000 \times 24.999 \times 24998 \times 24997 \times 24996 \; (= 9.8e+21)$$

Question 2 (6 points)

On scanne la séquence d'un génome bactérien de 4 Mb pour y trouver toutes les occurrences du motif TACGATGC, en acceptant au plus 2 substitutions. On considère que les nucléotides sont répartis de façon indépendante, et qu'ils sont équiprobables dans ce génome. Comment calcule-t-on la probabilité d'observer une occurrence du motif à une position donnée du génome ?

a. Expliquez le raisonnement qui permet de déterminer la distribution théorique de probabilité à utiliser.

Sur base des présupposés de l'énoncé, on peut modéliser l'occurrence du motif à une position donnée du génome sur base d'un modèle de Bernoulli généralisé : chaque position du motif est considéré comme un essai avec 4 événements possibles (A, C, G, T) étant chacun associé à une probabilité fixe ($P_A = P_C = P_G = P_T = 0.25$). On teste successivement la correspondance entre les nucléotides successifs du motif et les nucléotides avec lesquels ils sont alignés sur le génome. Cette comparaison peut donner soit un succès (nucléotides identiques) soit un échec (nucléotides différents). Il y a donc $n=9$ essais, et on suppose qu'ils sont indépendants et que la probabilité de succès est constante ($p=0.25$).

Dans le cadre d'un tel modèle de Bernoulli, on peut mesurer la probabilité d'observer une occurrence avec au plus $m=2$ substitutions (ou, de façon équivalente, avec au moins $x = n - m = 9 - 2 = 7$ identités) au moyen de la loi binomiale.

b. Ecrivez la formule de cette distribution théorique en expliquant (dans le contexte général) ce que représente chaque symbole.

La densité binomiale indique la probabilité d'observer exactement x identités entre le motif et un segment génomique de 9 nucléotides.

$$P(X = x) = C_n^x p^x (1-p)^{n-x} = \frac{n!}{x!\,(n-x)!} p^x (1-p)^{n-x}$$

où

- $p=0.25$ est la probabilité de trouver une identité à une position donnée du motif
- $n=9$ est le la taille du motif
- X est une variable aléatoire représentant le nombre d'identités
- x est une valeur particulière pour le nombre d'identités observées (nous considérerons successivement $x=7$, $x=8$ et $x=9$).

La probabilité d'observer au plus 2 substitutions (ou de façon équivalente, au moins 7 identités) se calcule par la somme des densités.

$$P(X \geq x) = \sum_{i=x}^{n} C_n^i p^i (1-p)^{n-i}$$

c. Ecrivez cette même formule en remplaçant les symboles par les valeurs numériques extraites de l'énoncé. Vous ne devez pas calculer le résultat final.

$$P(X \geq 7) = \sum_{i=7}^{9} C_9^i 0.25^i (1-0.25)^{9-i}$$

d. Discutez de la pertinence des hypothèses de travail.

Le modèle de Bernoulli est contestable pour plusieurs raisons.

L'hypothèse d'indépendance entre nucléotides successifs ne correspond pas à la réalité des séquences génomiques, où l'on observe de fortes dépendances locales (par exemple fréquence élevée de séquences poly-A et poly-T dans les séquences non-codantes, dépendance entre fréquences nucléotidiques et phase de codon dans les séquences codantes, évitement du dinucléotide CpG dans les génomes de mammifères, …).

L'équiprobabilité entre résidu peut s'observer dans certains cas, mais en général il existe une disparité entre fréquences de A+T et C+G dans les séquences génomiques. De plus ces fréquences dépendent du contexte génomique (régions intergéniques, promoteurs, introns, exons codants ou non-codants, …).

Question 3 (10 points)

On veut évaluer l'impact d'un médicament sur la concentration d'une protéine dans une lignée cellulaire, en mesurant la fluorescence d'un marqueur associé. On suppose que les fluctuations expérimentales suivent une loi normale. Les mesures sont indiquées ci-dessous.

Cellules traitées : 92 96 102 86 118 110 94 102

Cellules non-traitées : 125 86 76 101 95 136 108 113

a. Calculez les moyennes, écarts-types et les médianes des deux échantillons.

Paramètres d'échantillons	Cellules traitées	Non-traitées
Moyenne	100	105
Ecart-type	9.64	18.55
Médiane	99	104.5

b. Pour chaque échantillon, estimez la moyenne et l'écart-type de la population.

Paramètres de populations estimés	Cellules traitées	Non-traitées
Moyenne	100	105
Ecart-type	10.31	19.83

c. Quel test choisiriez-vous pour évaluer si le traitement a un effet ou non ? Justifiez vos choix en vous basant sur l'énoncé, et sur les résultats de la sous-question b.

On postule que les données suivent une distribution normale, on peut donc appliquer un test paramétrique. Il faut dès lors évaluer si on peut ou non postuler l'homoscédasticité (autrement dit, supposer que les populations dont ces échantillons sont extraits ont des variances égales). Les écarts-types des échantillons semblent assez différents. Idéalement on appliquerait un test d'égalité des variances (non vu au cours), mais la taille des échantillons est trop faible pour que ce test soit fiable. Nous pouvons donc adopter une attitude raisonnable en partant d'une hypothèse de travail d'hétéroscadicité (on suppose que les populations ont des variances différentes). On va donc appliquer un test de Welch.

Comme l'énoncé ne fournit aucune indication concernant le sens du test, on applique un test bilatéral.

a. Ecrivez la formule de l'hypothèse nulle et décrivez-la en une phrase.

$$H_0: \mu 1 = \mu 2$$
$$H_1: \mu 1 \neq \mu 2$$

L'hypothèse nulle est que les populations dont les échantillons sont extraits ont des moyennes égales. L'hypothèse alternative est que ces populations ont des moyennes différentes, quel que soit le sens de cette différence.

b. Ecrivez la formule de la statistique du test choisi (note : j'ai corrigé une erreur du formulaire : dans la formule du test de Welch il on utilise les écarts-types estimés pour les populations et non les écarts-types d'échantillons).

$$t_W = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$$

c. Réécrivez cette formule en remplaçant chaque symbole par la valeur numérique correspondante.

$$t_W = \frac{100 - 105}{\sqrt{\frac{10.31^2}{8} + \frac{19.83^2}{8}}}$$

d. Calculez la valeur de la statistique de test.

$$t_W = -0.63$$

e. Calculez le nombre de degrés de liberté (*d.l.*).

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}} = \frac{\left(\frac{9.64^2}{8} + \frac{18.55\text{^}2}{8}\right)^2}{\frac{9.64^4}{8^27} + \frac{18.55^4}{8^27)}} = 10.53$$

f. Indiquez la p-valeur

Dans la table *t* (fournie à l'examen) les degrés de liberté sont entiers. Le test de Welch se base sur des nombres rationnels pour les degrés de libertés, mais on peut interpoler la valeur entre les lignes

correspondant à $\nu = 10$ et $\nu = 11$. Avec une statistique t_W=*-0.63* la p-valeur d'un test bilatéral est approximativement $p \sim 0.5$, voire un peu plus élevée (note : la p-valeur précise, calculée avec R, vaut *p=0.54*).

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$
one-tail	0.50	0.25	0.20
two-tails	1.00	0.50	0.40
df			
1	0.000	1.000	1.376
2	0.000	0.816	1.061
3	0.000	0.765	0.978
4	0.000	0.741	0.941
5	0.000	0.727	0.920
6	0.000	0.718	0.906
7	0.000	0.711	0.896
8	0.000	0.706	0.889
9	0.000	0.703	0.883
10	0.000	0.700	0.879
11	0.000	0.697	0.876
12	0.000	0.695	0.873

g. Expliquez en une phrase ou deux comment on interprète cette p-valeur.

Cette p-valeur représente la probabilité d'obtenir une statistique *t* au moins aussi importante (en valeur absolue) sous hypothèse nulle (c'est-à-dire, si on avait tiré des échantillons de populations ayant des moyennes égales).

h. Quelle décision prenez-vous à l'issue du test ? Justifiez.

La p-valeur est très élevée, on rejette donc l'hypothèse nulle.