PROBABILITES ET STATISTIQUES POUR LA BIOLOGIE
(STAT1, ENSBBAU16L) – EXAMEN – 8 JANVIER 2018

Calculatrices Autorisées ; Documents Non Autorisés.

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

Question 1 (4 points)

On a mesuré les niveaux d'expression des 25.000 gènes d'un organisme donné, et on sélectionne les 27 gènes les plus fortement exprimés. Combien de possibilités distinctes existe-t-il pour l'ensemble de gènes sélectionnés, si l'on ne tient pas compte de l'ordre parmi ces 27 gènes ? 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 non-ordonné sans remise, on utilise donc le coefficient binomial (choix de *x=27* gènes parmi *n=25000*).

$$C_n^x = \frac{n!}{x! \cdot (n-x)!} = C_{25.000}^{27} = \frac{25.000!}{27! \cdot 24.973!} = 5{\times}10^{90}$$

Question 2 (6 points)

On a établi que motif de liaison d'un facteur transcriptionnel peut être représenté par le consensus STGAWNNNWTCAS, où S signifie "C ou G", W "A ou T" et N n'importe quel nucléotide. On désire calculer le nombre attendu d'occurrences du motif dans deux génomes.

a. Un génome bactérien de 5 Mb, dont les fréquences nucléotidiques sont approximativement égales.
b. Un génome de levure de 12 Mb, dont les fréquences moyennes de nucléotides sont F(A) = F(T) =0.3; F(G) = F(C) = 0.2.

Expliquez les hypothèses de travail, la procédure, les détails des calculs et les résultats pour chacun des deux génomes. A chaque étape, justifiez vos choix en explicitant les modèles et de calculs de probabilités.

Hypothèses de travail: modèle de Bernoulli, avec nucléotides indépendants. Dans le cas de la bactérie, mais pas celui de la levure, on supposera les nucléotides équiprobables.

Procédure :

a. Calcul de la probabilité d'occurrence du motif à une position génomique, sur base d'un modèle de Bernoulli.
b. Calcul du nombre d'occurrences attendues.

Probabilité d'occurrences :

- Les probabilités a priori des nucléotides A, C, G, T sont fournies dans l'énoncé.
- La probabilité des nucléotides incomplètement spécifiés (S, W, N) est la somme des probabilités des nucléotides qu'ils représentent, car ces nucléotides sont mutuellement incompatible (on ne peut aps trouver deux nucléotides différents à la même position).
    - Pour la bactérie ceci donne :

$$P(S) = P(C) + P(G) = 0.25 + 0.25 = 0.5$$
$$P(W) = P(A) + P(T) = 0.25 + 0.25 = 0.5$$

    - Pour la levure :

$$P(S) = P(C) + P(G) = 0.2 + 0.2 = 0.4$$
$$P(W) = P(A) + P(T) = 0.3 + 0.3 = 0.6$$

- Selon un modèle de Bernoulli, la probabilité d'une séquence est le produit des probabilités de ses résidus (en vertu de l'indépendance).
    - Pour la bactérie ceci donne :

$$P(\text{STGAWNNNWTCAS}) = P(S) \cdot P(T) \cdot P(G) \ldots P(C) \cdot P(A) \cdot P(S)$$
$$= P(S)^2 \cdot P(T)^2 \cdot P(G) \cdot P(A)^2 \cdot P(W)^2 \cdot P(C) \cdot P(N)^3 = 0.5^4 \cdot 0.25^6 \cdot 1^3$$
$$= 1.52 \times 10^{-5}$$

- Pour la levure ceci donne :

$$P(\text{STGAWNNNNWTCAS}) = P(S) \cdot P(T) \cdot P(G) \ldots P(C) \cdot P(A) \cdot P(S)$$
$$= P(S)^2 \cdot P(T)^2 \cdot P(G) \cdot P(A)^2 \cdot P(W)^2 \cdot P(C) \cdot P(N)^3$$
$$= 0.4^2 \cdot 0.3^2 \cdot 0.2 \cdot 0.3^2 \cdot 0.6^2 \cdot 0.2 \cdot 1^3 = 1.86 \times 10^{-5}$$

Nombre d'occurrences attendues.

Le motif est réverse complémentaire palindromique. Il suffit donc de le chercher sur un seul brin pour trouver toutes ses occurrences distinctes (l'occurrence trouvée à la même position sur l'autre brin n'est pas prise en compte, comme indiqué lors de l'examen).

Pour la bactérie

$$G = 5 \times 10^6$$
$$E(\text{STGAWNNNNWTCAS}) = G \cdot P(\text{STGAWNNNNWTCAS}) = 5 \times 10^6 \cdot 1.52 \times 10^{-5} = 76.3$$

Pour la levure

$$G = 12 \times 10^6$$
$$E(\text{STGAWNNNNWTCAS}) = G \cdot P(\text{STGAWNNNNWTCAS}) = 12 \times 10^6 \cdot 1.86 \times 10^{-5} = 223.9$$

Question 3 (4 points)

a. Expliquez les problèmes que peut éventuellement poser la variance d'échantillon pour estimer la dispersion d'une population.

- La variance d'échantillon est un estimateur biaisé de la variance de la population. Il existe en effet un biais systématique: la variance d'échantillon sous-estime la variance de la population, car elle calcule les carrés des distances entre chaque mesure et la moyenne d'échantillon plutôt que la moyenne de population (qui est inconnue). Or, la moyenne d'échantillon est par définition la valeur qui minimise le carré des distances, la variance d'échantillon est donc une borne inférieure pour la variance réelle de la population.
- La variance est également particulièrement sensible à la présence de valeurs aberrantes, dont les différences (excessives) interviennent au carré.

b. Que fait-on pour obtenir un estimateur non-biaisé de l'écart-type de population ?

Le biais systématique peut être corrigé simplement en remplaçant la taille d'échantillon *n* par *n-1* dans la formule de la variance.

c. Citez un estimateur de dispersion réputé robuste, qui n'est pas basé sur la variance.

L'espace interquartile est un estimateur robuste aux valeurs aberrantes. Cependant, il n'est valide que pour des échantillons de taille élevée, car il est sensible aux fluctuations des valeurs pour des petits échantillons, et converge moins rapidement que l'écart-type.

d. Calculez ces statistiques pour l'échantillon suivant, en fournissant le détail du calcul : $x = \{2, 6, 4, 2, 8, 4, 8, 30\}$.

Effectif d'échantillon	n	*8*
Moyenne d'échantillon	$\bar{x}$	*8*
Variance d'échantillon	s^2	*74*
Ecart-type d'échantillon	s	*8.6*
Premier quartile	*Q1*	*3.5*
Troisième quartile	*Q3*	*8*
Espace interquartile	*IQR = Q3 - Q1*	*4.5*
Estimateur non biaisé de la variance	$\hat{\sigma}^2$	*84.6*
Estimateur non biaisé de l'écart-type	$\hat{\sigma}$	*9.2*

Question 4 (6 points)

On effectue une analyse d'expression différentielle pour les 5000 gènes d'une bactérie au moyen de biopuces transcriptomiques, avec 3 répliques par condition (traité *versus* contrôle). Pour un gène donné, on observe un niveau moyen de 15 pour les échantillons soumis au traitement, 5 pour les contrôles, avec un écart-type de 5 dans les deux groupes. Sur base de données publiées précédemment, on postule que les fluctuations d'expression de ce gène suivent une distribution approximativement normale pour chacun des deux groupes.

a. Sur quel critère se base votre choix pour le test de comparaison de moyenne ? Justifiez les différents éléments de votre choix.

On postule que les données d'expression suivent une distribution normale, on peut donc appliquer un test paramétrique. Comme les écarts-types des échantillons sont égaux, on peut se baser sur une hypothèse de travail d'homoscédasticité (autrement dit, postuler que les populations dont ces échantillons sont extraits ont des variances égales). On peut donc utiliser un test de Student.

A priori, on s'intéresse aux effets du traitement dans les deux sens (sur-expression ou sous-expression). On va donc effectuer un test bilatéral.

b. 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.

c. Ecrivez la formule de la statistique du test choisi, en indiquant la valeur de chaque symbole, puis le résultat final.

$$t_S = \frac{\hat{\delta}}{\hat{\sigma}_\delta} = \frac{\overline{x}_2 - \overline{x}_1}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$\bar{x}_1 = 15; \bar{x}_x = 5; n_1 = n_2 = 3; s_1 = s_2 = 5$$
$$t_S = 2$$

d. Indiquez la p-valeur et fournissez-en une interprétation en une phrase.

$$p = 0.116$$

e. En tenant compte du nombre de gènes analysés, quelle serait la e-valeur associée à ce gène ? Interprétez cette e-valeur en une phrase.

$$E = p * N = 0.12 * 5000 = 580$$

f. Quels sont les paramètres qui expliquent le lien entre le résultat du test et la différence apparemment élevée d'expression (rapport de 3 entre les échantillons traités et les contrôles) ?

La différence entre les moyennes d'échantillons semble intuitivement élevée : l'expression est 3 fois plus élevée pour les traités (m1 = 15) que pour les contrôles (m2 = 5), et la différence (m2 – m1 = 10) vaut 2 fois l'écart-type d'échantillon (s1 = s2 = 5). Cependant, la formule du t-test prend en compte les effectifs des échantillons, qui sont très faibles (n1 = n2 = 3). La valeur de la statistique du test (t = 2) est donc finalement assez faible.

En pratique, ceci indique qu'avec 3 échantillons par groupe les moyennes d'échantillon ne fournissent pas d'estimation fiable des moyennes de populations, et on ne dispose pas d'une puissance suffisante pour pouvoir rejeter l'hypothèse nulle. Il faudrait recommander aux biologistes de refaire l'expérience avec des échantillons plus importants.