

**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.

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.

Question 3 (4 points)

- a. Expliquez les problèmes que peut éventuellement poser la variance d'échantillon pour estimer la dispersion d'une population.
- b. Que fait-on pour obtenir un estimateur non-biaisé de l'écart-type de population ?
- c. Citez un estimateur de dispersion réputé robuste, qui n'est pas basé sur la variance.
- d. Calculez ces statistiques pour l'échantillon suivant, en fournissant le détail du calcul :
 $x = \{2, 6, 4, 2, 8, 4, 8, 30\}$.

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 échantillons 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.
- b. Ecrivez la formule de l'hypothèse nulle et décrivez-la en une phrase.
- c. Ecrivez la formule de la statistique du test choisi, en indiquant la valeur de chaque symbole, puis le résultat final.
- d. Indiquez la p-valeur et fournissez-en une interprétation en une phrase.
- 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.
- 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) ?

Formules de probabilités et statistique

Probabilités et statistique pour la biologie (STAT1)

Jacques van Helden

2018-01-08

Combinatoire

Nom	Conditions	Formule
Permutations (factorielle)		$n! = n \cdot (n-1) \cdot \dots \cdot 2 \cdot 1$
Arrangements	Sans remise, ordonné	$A_n^x = \frac{n!}{(n-x)!} =$ $n \cdot (n-1) \cdot \dots \cdot (n-x+1)$
Combinaison (<i>choose</i> , <i>coefficient binomial</i>)	Sans remise, sans ordre	$\binom{n}{x} = C_n^x = \frac{n!}{x!(n-x)!}$

Concepts de probabilité

Description	Conditions	Formule
Définition fréquentielle de la probabilité		$P(A) = \lim_{n \rightarrow \infty} \frac{n_A}{n}$
Probabilité de non-réalisation		$P(\neg A) = 1 - P(A)$
Probabilités conditionnelles		$P(A B) = \frac{P(A \wedge B)}{P(B)}$ $P(B A) = \frac{P(A \wedge B)}{P(A)}$
Probabilité de A ou B	En général	$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
Probabilité de A ou B	Evénements mutuellement exclusifs	$P(A \vee B) = P(A) + P(B)$
Probabilité de A et B	En général	$P(A \wedge B) = P(A) \cdot P(B A)$
Probabilité de A et B	Evénements indépendants	$P(A \wedge B) = P(A) \cdot P(B)$
Règle de Bayes		$P(A \wedge B) = P(A) \cdot P(B A) = P(B) \cdot P(A B)$ $\implies P(A B) = \frac{P(A) \cdot P(B A)}{P(B)}$ $\implies P(B A) = \frac{P(B) \cdot P(A B)}{P(A)}$

Distributions de probabilité discrètes

Géométrique

- Conditions : nombre d'échecs avant le premier succès dans un schéma de Bernoulli
- Densité :

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

- Répartition :

$$P(X \leq x) = \sum_{i=0}^x (1 - p)^i p$$

- Moyenne : $\mu_G = (1 - p)/p$
- Variance : $\sigma_G^2 = \frac{(1-p)}{p^2}$

Binomiale

- Conditions : Nombre de succès au cours d'une série d'essais indépendants avec probabilité constante de succès (Schéma de Bernoulli)
- Densité :

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

- Répartition :

$$P(X \leq x) = \sum_{i=0}^x C_n^i p^i (1 - p)^{n-i}$$

- Moyenne : $\mu_B = np$
- Variance : $\sigma_B^2 = np(1 - p)$
- Rapport moyenne/variance: $\sigma_B^2 < \mu_B$

Poisson

- Conditions : nombre de succès observés au cours d'un intervalle de temps, en fonction du nombre attendu (λ)
- Application : approximation de la binomiale quand $n \rightarrow \infty, p \rightarrow 0$ et $\mu = np$ faible ($\mu_B \rightarrow \lambda$)
- Densité :

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

- Répartition :

$$P(X \leq x) = \sum_{i=0}^x \frac{e^{-\lambda} \lambda^i}{i!}$$

- Moyenne : $\mu_P = \lambda$
- Variance : $\sigma_P^2 = \lambda$
- Rapport moyenne/variance: $\sigma_P^2 = \mu_P$

Hypergéométrique

- Conditions : Tirage non ordonné, sans remise dans un ensemble fini avec deux catégories.
- Exemple-type: urne avec boules de deux couleurs
- Densité :

$$P(X = x) = \frac{C_m^x C_n^{k-x}}{C_{m+n}^k}$$

- Répartition :

$$P(X \leq x) = \sum_{i=x}^{\min(k,m)} \frac{C_m^i C_n^{k-i}}{C_{m+n}^k}$$

- Moyenne : $\mu_H = k \cdot \frac{m}{m+n}$
- Variance : $\sigma_H^2 = \frac{k \frac{m}{N} (1 - \frac{m}{N})(N-k)}{(N-1)}$; $N = m + n$

Echantillonnage et estimation

- Les symboles grecs (μ, σ) correspondent aux statistiques de population, les symboles romains (\bar{x}, s) aux statistiques d'échantillon.
- L'accent circonflexe ($\hat{\cdot}$) indique les estimateurs de paramètres de population calculés à partir de paramètres d'échantillons.

Symbole	Description
N	Taille (nombre d'individus) de la population.
$\mu = \frac{1}{N} \sum_{i=1}^N x_i$	Moyenne de la population.
$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$	Variance de la population
$\sigma = \sqrt{\sigma^2}$	Écart-type de la population
n	Effectif (nombre d'individus) de l'échantillon.
$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$	Moyenne d'échantillon.
$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$	Variance de l'échantillon
$s = \sqrt{s^2}$	Écart-type de l'échantillon
$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$	Estimateur non-biaisé de la variance de la population.
$\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$	Estimateur non-biaisé de l'écart-type de la population.
$\langle \sigma_{\bar{X}} \rangle = \frac{\hat{\sigma}}{\sqrt{n}}$	Erreur standard: écart-type attendu sur la moyenne d'échantillon.
$\bar{x} \pm \frac{\hat{\sigma}}{\sqrt{(n)}} \cdot t_{1-\alpha/2}^{n-1}$	Intervalle de confiance autour de la moyenne.

Test de comparaison de moyennes

Symbole	Description
μ_1, μ_2	Moyennes respectives des populations 1 et 2.
σ_1, σ_2	Écart-types respectifs des populations 1 et 2.
n_1, n_2	Effectifs (nombre d'individus) des échantillons prélevés sur les populations 1 et 2.
\bar{x}_1, \bar{x}_2	Moyennes d'échantillons.
$\delta = \mu_2 - \mu_1$	Différence entre les moyennes des populations.
$d = \hat{\delta} = \hat{\mu}_2 - \hat{\mu}_1 = \bar{x}_2 - \bar{x}_1$	$d =$ Taille d'effet : dans un test de comparaison de moyennes, il s'agit de la différence entre les moyennes d'échantillons, utilisée comme estimateur de δ .
s_1^2, s_2^2	Variances mesurées sur les échantillons.
$\hat{\sigma}_p = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}$	Écart-type groupé (<i>pooled standard deviation</i>), utilisé comme estimateur de l'écart-type commun des deux populations, en supposant leurs variances égales (hypothèse de travail d'homoscédasticité).
$\hat{\sigma}_\delta = \hat{\sigma}_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$	Erreur standard sur la différence entre moyennes, en supposant que les populations ont la même variance (test de Student).
$t_S = \frac{\hat{\delta}}{\hat{\sigma}_\delta} = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$	Statistique t de Student
$t_W = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	Statistique t de Welch

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
Z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										