Probabilités et statistiques pour la biologie (STAT1, ENSBBAU16L) – examen de rattrapage – 11 MAI 2018

Probabilités et statistique pour la biologie (STAT1)

Jacques van Helden

11 mai 2018

Contents

Quesiton 1 (4 points)

```
p <- 1/6 ## Probability of success (getting a 6)
n <- 60 ## Number of trials
x <- 20

exp <- p * n

pval <- pbinom(q = x-1, size = n, prob = p, lower.tail = FALSE)

## Note: for the exercise, we can explicitly compute the p-value
## from the raw binomial formula
i <- x:n
pval2 <- sum(choose(n = n, k = i) * p^i *(1-p)^(n-i))
```

On effectue une série de 60 tirs de dés, et on observe que la valeur 6 tombe 20 fois.

a. *En supposant le dé équilibré, combien de 6 se serait-on attendu à obtenir ?*

$$< x >= p \cdot n = 0.167 \cdot 60 = 10$$

b. *Comment pourrait-on calculer la p-valeur du résultat obtenu ? Justifiez le choix d'une distribution théorique de probabilité.*

Chaque lancer de dé peut donner lieu à 6 résultats possibles (les 6 faces du dé), et on s'intéresse à un événement particulier (la valeur 6). La série de lancers correspond à un schéma de Bernoulli : chaque lancer (essai) peut donner lieu soit à un succès (tirage du 6) ou un échec (les 5 autres faces), les essais successifs sont indépendants et ont une probabilité de succès constante ($p = 1/6$). On peut donc utiliser la **distribution binomiale**.

c. *Ecrivez la formule générique (avec les symboles) pour calculer la p-valeur, puis remplacez les symboles par les valeurs de paramètres extraites de l'énoncé. Il n'est pas nécessaire de calculer le résultat final.*

$$P = \sum_{i=x}^{n} C_i^n \cdot p^i \cdot (1-p)^{n-i}$$

$$P = \sum_{i=20}^{60} C_i^{60} \cdot \frac{1}{6}^i \cdot \frac{5}{6}^{60-i} = 0.00123$$

d. *Après calcul (avec un ordinateur), on obtient une p-valeur de 0.001. Comment interprétez-vous ce résultat ?*

La p-valeur représente la probabilité d'obtenir un résultat au moins aussi extrême que celui obtenu (autrement dit, au moins 20 tirages d'un 6 sur 60 essais) sous hypothèse nulle (dans ce cas, l'hypothèse nulle est que les 6 faces sont équiprobables).

Une p-valeur de $P = 0.001$ indique que la probabilité d'obtenir au moins 20 fois le 6 sur 60 tirages est très faible, il est donc vraisemblable que le dé soit pipé.

Question 2 (6 points)

```r
G <- 1e+6 # Genome size
motif <- "ATACGHNWKATGC"
k <- nchar(motif)

## Prior residue probabilities
prior <- c("A" = 0.3, "T" = 0.3, "C" = 0.2, "G" = 0.2)
prior["N"] <- 1
prior["W"] <- prior["A"] + prior["T"]
prior["H"] <- 1 -  prior["G"]
prior["K"] <- prior["T"] + prior["G"]

## Probability to find an instance of the motif at a given position
p <- prod(prior[unlist(strsplit(motif, ""))])

E <- 2 * p * G
```

On scanne la séquence d'un génome bactérien de 1 Mb sur les deux brins, pour y trouver toutes les occurrences exactes du motif $ATACGHNWKATGC$ (voir table IUPAC ci-dessous pour les codes des nucléotides ambigus). On suppose que les nucléotides du génome se succèdent de façon indépendante. La composition nucléotidique du génome est de 30% de A et T et 20% de C et G.

Table : IUPAC ambiguous nucleotide code

IUPAC	Nucleotides	Mnemo
A	A	Adenine
C	C	Cytosine
G	G	Guanine
T	T	Thymine
R	A or G	puRine
Y	C or T	pYrimidine
W	A or T	Weak hydrogen bonding
S	G or C	Strong hydrogen bonding
M	A or C	aMino group at common position
K	G or T	Keto group at common position

IUPAC	Nucleotides	Mnemo
H	A, C or T	not G
B	G, C or T	not A
V	G, A, C	not T
D	G, A or T	not C
N	G, A, C or T	aNy

a. *Comment calcule-t-on la probabilité des résidus dégénérés (K, W, H, N) ? Indiquez la formule générique en termes de probabilités de résidus ($P_A, P_C, \ldots, P_K, P_N$), puis remplacez les symboles par des valeurs, et calculez le résultat. Justifiez votre choix de la formule utilisée.*

A chaque position d'une séquence on peut observer 4 événements complémentaires : A, C, G, ou T. Les résidus dégénérés consistent en unions de ces événements. Comme ils sont mutuellement exclusifs (on ne peut pas observer duex résidus distincts à la même position) la probabilité de leur union est la somme des probabilités.

$$P_K = P_T + P_G = 0.5$$
$$P_W = P_T + P_A = 0.6$$
$$P_H = P_A + P_C + P_T = 1 - P_G = 0.8$$
$$P_N = P_A + P_C + P_G + P_T = 1$$

b. *Quelle est la probabilité de trouver une occurrence du motif à une position donnée du génome ? Expliquez les étapes de votre raisonnement, en justifiant les modèles probabilistes.*

Pour trouver une occurrence du motif à une position donnée du génome, il faut trouver trouver le premier résidu à la première position aligné (premier événement), et le second résidu à la seconde (second événement), et ainsi de suite jusqu'à la dernière position. Puisqu'on suppose que les nucléotides successifs sont indépendants, la probabilité jointe de cet ensemble d'événements est le produit de leurs probabilités respectives.

$$P(ATACGHNWKATGC) = P(A)\cdot P(T)\cdot P(A)\cdot P(C)\cdot P(G)\cdot P(H)\cdot P(N)\cdot P(W)\cdot P(K)\cdot P(A)\cdot P(T)\cdot P(G)\cdot P(C) = 0.3\cdot 0.3\cdot 0.3\cdot 0$$

c. *Quel est le nombre d'occurrences attendues au hasard sur l'ensemble du génome ? Expliquez le raisonnement, en justifiant le choix du modèle de calcul.*

Le nombre attendu d'occurrences est obtenu en multipliant la probabilité d'occurrences par le nombre de positions envisagées. Comme on scanne les deux brins, ce nombre de positions est 2 fois la taille du génome. Ceci respose sur l'hypothèse de travail (simplificatrice) selon laquelle les positions successives sont indépendantes.

$$E = p \cdot N = p \cdot 2 \cdot G = 9.33 \times 10^{-7} \cdot 2 \cdot 10^6 = 1.87$$

Question 3 (10 points)

*Un chercheur a mesuré par qPCR le niveau d'expression d'un gène d'intérêt à partir d'échantillons sanguins prélevés chez 50 patients ($n_p = 50$) et chez 50 sujets témoins ($n_t = 50$). 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.*

```
mp <- 21
mt <- 10
np <- nt <- n <- 50
sp <- st <- s <- 15

t <- (mp - mt) / sqrt((np * sp^2 + nt * st^2) / (np + nt -2) *(1/np + 1/nt))

## Simplified formula for equal sample sizes
t1 <- (mp - mt) / sqrt(n * s^2 / (n -1) *(2/n))

nu <- np + nt -2

pval <- 2*pt(q = t, df = nu, lower.tail = FALSE)
```

a. *Le choix du test de Student vous semble-t-il approprié ? Justifiez le choix du chercheur.*

   La première hypothèse de travail du test de Student est la normalité des deux populations dont les échantillon sont extraits. 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 d'éventuels 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 >30). La taille de nos échantillons ($n_1 = n_2 = 50$) justifie donc le recours à un test paramétrique.

   La seconde hypothèse de travail du test de Student est que les populations d'où proviennent les échantillons ont la même variance (homoscédasticité). 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.

b. *Quelles auraient été les situations alternatives possibles, et quels tests auraient été appropriés ?*

   Si les variances d'échantillons avaient montré de fortes disparités, nous aurions recouru au test de Welch plutôt quà celui de Student.

   Si les échantillons avaient été de plus petite taille, en absence d'indications de la normalité des populations, nous aurions recouru à un test non-paramétrique.

c. *Sachant qu'a priori on ignore dans quel sens la maladie pourrait affecter le niveau d'expression de ce gène, formulez l'hypothèse nulle et expliquez-la en une phrase.*

   On effectue un test bilatéral. L'hypothèse nulle est que les populations ont la même moyenne. 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 $21 \neq 10$).

$$H_0 : \mu_p = \mu_t$$
$$H_A : \mu_p \neq \mu_t$$

d. *Sur base du formulaire joint, calculez la statistique t de Student.*

$$t_S = \frac{\hat{\delta}}{\hat{\sigma}_\delta} = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{21-10}{\sqrt{\frac{50\cdot 15^2+50\cdot 15^2}{15+15-2}\left(\frac{1}{50}+\frac{1}{50}\right)}} = 3.63$$

e. *Indiquez, en vous basant sur la table t ci-jointe, la p-valeur correspondante.*

Le nombre de degrés de liberté vaut $\nu = n_p + n_t - 2 = 98$.

Avec 100 degrés de liberté, la valeur de la statistique t la plus élevée dans la table de Student est $t = 3.291$, qui correspond à une p-valeur $P = 0.001$. Nous pouvons donc conclure que la p-valeur correspondant à une statistique t de 3.667 est inférieure à 0.001.

Pour la confirmation, la p-valeur calculée pour une statistique $t = 3.63$ avec $\nu = 98$ degrés de liberté vaut $P = 4.53 \times 10^{-4}$.

f. *Interprétez la p-valeur, et aidez le chercheur à tirer les conclusions concernant l'impact éventuel de la maladie sur l'expression de ce gène.*

La P-valeur d'un test d'hypothèse indique la probabilité d'obtenir un résultat au moins aussi extrême que la statistique observée 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.

$P = P((T \leq -3.63) \vee (T \geq 3.63)) = 2 \cdot P(T \geq 3.63)$

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

La p-valeur obtenue ($P = 4.53 \times 10^{-4}$) est très faible, et on peut donc rejeter l'hypothèse nulle.