Examen de biostatistiques – L3 MIV

M. Bailly-Bechet & H. Haned

8 juin 2010

Documents autorisés. Échanges interdits. Calculatrices inutiles mais autorisées. Bonne humeur préférable. Durée de l'épreuve : 2h00.

Cette épreuve est divisée en trois parties indépendantes. Les résultats proposés peuvent être employés même s'ils n'ont pas été démontrés.

1 Un exercice sans rapport avec les rayons Γ

La fonction Γ (prononcer gamma) est une fonction définie par une intégrale, comme suit :

$$\Gamma(k) = \int_0^\infty e^{-\lambda t} \lambda^k t^{k-1} dt. \tag{1}$$

Elle n'a pas de forme analytique simple. On la présente souvent comme une généralisation de la fonction factorielle à l'ensemble des réels.

1. À l'aide d'une intégration par parties, montrez que $\Gamma(k+1) = k\Gamma(k)$.

Il existe une loi de probabilité, dite loi gamma, qui est construite à partir de cette fonction. C'est une loi à deux paramètres (comme la loi normale), λ et k. Sa densité de probabilité est :

$$p_\Gamma(x|\lambda, k) = \frac{x^{k-1}\lambda^k e^{-\lambda x}}{\Gamma(k)}. \tag{2}$$

On suppose que la variable aléatoire X est une v.a. continue qui suit une loi Γ de paramètres λ et k.

2. Vérifiez que la formule 2 correspond bien à celle d'une densité de probabilité.

3. On rappelle que l'espérance d'une v.a. Z de loi continue $p(Z = z)$ se calcule comme $\mathbb{E}(Z) = \int_{-\infty}^{\infty} zp(z)dz$. Montrez que l'on peut écrire l'espérance de X comme $\mathbb{E}(X) = \frac{\Gamma(k+1)}{\lambda\Gamma(k)}$. Déduisez-en l'espérance de X en fonction de k et λ.
4. Par le même raisonnement, calculez la variance $\mathbb{V}(X)$ en fonction de k et λ.

2 Il est lourd avec ses blagues...

Des généticiens ont caractérisé l'allèle présente dans chaque individu pour un gène qui existe en deux formes seulement, A et B, au sein d'un échantillon de la population française de 200 individus. 100 individus possèdent l'allèle A, et 100 l'allèle B. On parlera ensuite d'échantillon et d'individus A, et d'échantillon et d'individus B sans distinction. Quand on trie l'ensemble de ces individus de manière croissante par rapport à leur poids, on obtient l'ordre suivant (les pointillés représentent respectivement des suites de A ou de B) :

	A	A	...	A	...	A	A	B	B	...	B	...	B	B	A	A	...	A	A
Rang	1	2	...	25	...	49	50	51	52	...	100	...	149	150	151	152	...	199	200
Taille	t_1	t_2	...	t_{25}	...	t_{49}	t_{50}	t_{51}	t_{52}	...	t_{100}	...	t_{149}	t_{150}	t_{151}	t_{152}	...	t_{199}	t_{200}

Les chercheurs veulent savoir si la présence de l'allèle A ou de l'allèle B a un effet sur le poids des individus. Ils vont pour cela comparer les poids dans les deux échantillons de diverses manières.

1. Sans calculs, au vu de l'ordre présenté ci-dessus, l'allèle a-t-il un effet sur le poids ? Détaillez.
2. Les chercheurs procèdent tout d'abord à un test du t de Student, pour vérifier si les poids moyens des deux échantillons sont égaux ($\bar{t_A} = \bar{t_B}$). Donnez deux arguments pour lesquels l'emploi de cet test est probablement illégitime dans ce cas précis.

Le test de Student précédent n'a pas permis de rejeter l'hypothèse nulle selon laquelle les poids moyens des deux échantillons sont égaux. Les chercheurs, sceptiques, veulent vérifier par un autre test leur hypothèse. Ils vont donc

procéder au test du U de Wilcoxon-White-Manney. On rappelle que celui-ci consiste à calculer la somme des rangs pour un échantillon, à la centrer et la réduire, et à comparer la valeur observée (dans le cas des grands échantillons) à une variable normale centrée réduite, selon la formule :

$$U = \frac{\sum_{i=1}^{m} R(B_i) - \frac{m(m+n+1)}{2}}{\sqrt{mn(m+n+1)}} \tag{3}$$

On rappelle que dans cette formule, m est la taille de l'échantillon considéré (ici B) et n celle de l'autre échantillon. $R(B_i)$ représente le rang du i-ème élément de l'échantillon B dans le classement total.

3. À l'aide du tableau présenté plus haut, et en remarquant que les $R(B_i)$ ont une structure particulière, montrez que U vaut 0 dans ce cas.
4. Concluez sur ce test. Votre conclusion était-elle prévisible au vu de la construction du test de Wilcoxon-White-Manney ?
5. Si vous aviez à caractériser la différence entre ces deux échantillons à l'aide d'un test, comment procéderiez-vous ? On suppose que vous avez accès aussi bien au classement des données en fonction du poids qu'aux poidss eux-mêmes.

3 Bonjour, je m'appelle *Saccharomyces* et j'ai arrêté de boire depuis 3 jours

Un ingénieur travaille sur la levure, *Saccharomyces cerevisiae*. Cet organisme unicellulaire, eucaryote, est employé par des industriels car, en dégradant des sucres, il rejette de l'alcool et va, dans de bonnes conditions, être un élément essentiel du processus de création de la bière. Les expériences menées par l'ingénieur consistent à introduire une mutation empêchant un gène de s'exprimer, tour à tour et indépendamment dans chacun des 6280 gènes de l'organisme. Il obtient donc, après ses expériences, 6280 boîtes de culture, chacune contenant des levures qui expriment tous leurs gènes sauf 1 – et ce pour tous les gènes. Aucune culture n'est intacte, et aucune ne contient des levures pour lesquelles plusieurs gènes ne peuvent s'exprimer. L'objectif de l'expérience est de déterminer quels gènes jouent un rôle sur la production d'alcool. La variable mesurée est, pour chaque culture i, la production d'alcool moyenne par gramme de levure pendant une heure, notée x_i. Toutes les cultures sont placées dans des conditions strictement identiques.

Les industriels savent que, sans modification génétique de la levure, la production d'alcool par gramme de levure et par heure suit une loi normale de moyenne μ et d'écart-type σ. Ils veulent réaliser un test pour comparer la production d'alcool de chaque culture à cette référence, pour trouver les gènes qui ont un rôle (positif ou négatif) sur la production d'alcool chez la levure. Les hypothèses de ce test, pour chaque culture i, sont :
H_0 : La modification du gène i n'a aucun effet sur la production d'alcool.
H_1 : La modification du gène i modifie la production d'alcool.

1. Quelle est la statistique à calculer pour comparer la production d'alcool dans la culture i avec cette production standard ?
2. Comment le calcul précédent est-il modifié si on dispose de j mesures indépendantes $x_{i,j}$ pour chaque culture ?

En choisissant un risque de première espèce α de 5% pour chacun des 6280 tests, l'ingénieur obtient 352 résultats significatifs, pour lesquels le test permet de dire que la production d'alcool de la culture est significativement différente, au risque de 5%, de celle de référence.

3. On suppose que la modification génétique dans la culture 17 n'a aucun effet sur la production d'alcool. Quel est la probabilité, en fonction de α, que le test donne néanmoins comme résultat qu'elle est significativement différente ?
4. On suppose que les cultures 17 et 582 n'ont aucun effet sur la production d'alcool. Quelle est la probabilité, en fonction de α, que au moins l'un des deux tests sur ces cultures permette de rejeter H_0 ?
5. Montrez que, dans le cas de N cultures pour lesquelles H_0 est vraie (aucun effet sur la production d'alcool), la manière de tester précédente en trouvera n pour lesquelles le test permettra de rejetter H_0, au risque α, avec n suivant une loi de probabilité :

$$p(n = k) = \binom{N}{k} \alpha^k (1 - \alpha)^{N-k} \tag{4}$$

6. On peut montrer, en développant l'expression précédente, que le nombre moyen de cas ou H_0 sera rejetée alors qu'elle est vraie est de αN. À combien de faux positifs (c.-à-d. de résultats significativement différents de H_0 par hasard) peut-on s'attendre parmi les 352 précédents ?
7. Quelle méthode préconerisiez-vous pour réduire le nombre de faux positifs dans ce type d'étude où l'on effectue N fois le même test ?