

TEST DE CHI-CARRE & TABLES DE CONTINGENCE

Veillez à avoir lu les sections relatives à cette séance dans le document disponible sur le site « Introduction aux séances de TP de biostatistique » avant de vous présenter au TP.

Veillez aussi à maîtriser les notions apprises au(x) TP(s) précédent(s).

Exercice 1

Un poulailler compte 10000 animaux, appartenant à 4 races différentes (Cornish (C), Leghorn (L), Rhode Island (RI) et Sussex (S)). On supposera en outre que la répartition des races s'établit comme suit : 30% de C, 30% de L, 20% de RI et 20% de S. On souhaite constituer un échantillon au hasard de 70 poules de ce poulailler (pour simplifier, on suppose qu'il y a remise).

Dans Excel, utilisez la fonction *ALEA()* combinée avec la fonction *SI(...)* pour simuler cet échantillonnage. La fonction *SI(condition;valeur si vrai;valeur si faux)* peut s'utiliser pour attribuer la valeur à une cellule en fonction d'une condition. La condition peut contenir une référence cellulaire. Ainsi, par exemple, la condition $A2 \leq 0,3$ sera vraie si la valeur contenue dans la cellule A2 est inférieure ou égale à 0,3 (30%) ; dans ce cas, la fonction *SI* retourne la valeur notée « valeur si vrai ». Notez que la valeur en question peut être une formule (qui serait alors évaluée pour fournir une valeur), du texte (entre guillemets) ou une valeur. Si la condition n'est pas vérifiée, alors le 3^e argument sera évalué. En recopiant cette formule dans une plage de 70 cellules, vous obtiendrez donc un échantillon de 70 poules. Établissez un comptage dans votre échantillon pour chacune de 4 races de poules grâce à la fonction *NB.SI(plage;"condition")*. L'argument "condition" de cette fonction peut être du type ">10", "58", "Leghorn", ... ou faire référence à une cellule qui contient le texte de la condition.

Cet échantillon correspond-il bien au poulailler c'est-à-dire est-il représentatif de la proportion réelle de chaque race au sein du poulailler ? Pour répondre à cette question, nous allons réaliser un test de chi-carré, qui compare les effectifs observés aux effectifs attendus.

Formulez l'hypothèse nulle. Si H_0 est vraie, quelles devraient être les effectifs attendus pour chacune des races de poules ? Détaillez le calcul du chi-carré et calculez la probabilité d'avoir cette répartition des effectifs observés sous l'hypothèse nulle (fonction *LOI.KHIDEUX.N(...)*). La fonction inverse *KHIDEUX.INVERSE(proba;ddl)* vous permet de trouver la valeur seuil de la table de chi-carré en fonction du nombre de degrés de liberté. Remarque : Une autre fonction permet de réaliser un test de chi-carré à partir de deux plages de cellules : la fonction *TEST.KHIDEUX(plage observée;plage attendue)* renseigne directement la probabilité obtenue. Elle devrait être identique à celle calculée à partir de la valeur statistique observée avec la fonction *LOI.KHIDEUX.N(...)*.

Dans R, créez un vecteur avec les 4 effectifs observés et utilisez la fonction *chisq.test(eff,p=proba)* où *eff* est le vecteur des effectifs observés et *proba* le vecteur des probabilités attendues sous l'hypothèse nulle. Votre résultat est-il identique à celui obtenu dans Excel ?

Si vous enregistrez votre fonction *chisq.test(...)* dans une variable, par exemple nommée *c2*, vous pourrez accéder à certains éléments de votre test de chi-carré comme les valeurs attendues en tapant *c2\$expected*. D'autres éléments de *c2* sont accessibles, voir *names(c2)* pour voir lesquels.

Exercice 2

La fréquence cardiaque au repos chez le cheval dépend-elle de la race ?

Pour tenter d'apporter une réponse à cette question, des chevaux appartenant à 3 races ont été sélectionnés au hasard, et leur fréquence cardiaque a été mesurée. Les chevaux sont classés selon la race et le niveau de leur fréquence cardiaque (en battements/minute). Les effectifs observés sont les suivants :

	≤ 60 batt/min	> 60 batt/min
Shetland	8	12
Frison	10	10
Pur-Sang	6	14

Reproduisez cette table dans R en utilisant la fonction *matrix(...)* déjà vue précédemment. La fonction *as.table(...)* permettra de transformer la matrice en une table de contingence. Appliquez ensuite le test de chi-carré et le test exact de Fisher (*fisher.test(...)*) sur cette table. Comparez les résultats entre eux et par rapport à votre hypothèse nulle de départ.

Essayez successivement les fonctions suivantes et observez leur résultat:

- *prop.table(table,1/2)* : affiche la table des probabilités (1 = pour les rangs et 2 = pour les colonnes)
- *margin.table(table,1/2)* : affiche les totaux marginaux (1 = pour les rangs et 2 = pour les colonnes)
- *summary(table)* : résumé de la table avec test de chi-carré

Exercice 3

Un laboratoire met au point un nouveau vaccin contre le coronavirus entérique du chat. Il effectue donc une étude pilote sur un nombre réduit d'animaux répartis en 2 groupes (Vacciné vs Non vacciné) afin de déterminer si le vaccin a bien un effet protecteur. Quatre semaines après l'administration du vaccin ou du placebo, tous les animaux sont exposés au virus et quatre semaines plus tard on teste les selles des chats pour voir s'ils sont encore porteurs du virus. Voici les résultats obtenus :

	Porteur	Sain
Non Vacciné	6	3
Vacciné	2	5

Dans Excel, calculez la probabilité de cette table et de toutes les autres tables possibles à partir des mêmes totaux marginaux, grâce à la fonction *LOI.HYPERGEOMETRIQUE.N(...)*. La somme de toutes les probabilités doit valoir 100% (tous les cas de figure seront représentés). Faites un histogramme de ces probabilités. *Remarque* : n'oubliez pas de classer les tables selon un certain ordre pour avoir une distribution graphiquement cohérente).

Dans R, effectuez le test exact de Fisher permettant de tester l'hypothèse nulle (H_0 : la vaccination ne diminue pas le portage viral). Pourquoi appliquer ce test plutôt qu'un test de chi-carré ? Enfin, tirez la conclusion adéquate sachant que le test doit être réalisé de manière unilatérale : pour cela mentionnez l'argument *alternative="greater"* dans la fonction du test. *Remarque* : le logiciel prend comme référence la valeur au croisement de la première ligne et de la première colonne. Ainsi, si vous appliquez le test sur un tableau reproduit identiquement au tableau présenté ci-dessus, l'argument est *alternative="greater"*. En revanche, si le tableau est par exemple produit de telle manière que c'est le nombre de non vaccinés sains qui correspond au croisement de la première ligne et colonne, alors on indiquera *alternative="less"*.

Exercice facultatif pour le brain storming à domicile

Exercice vivement conseillé en entraînement : Sachant que le poids des bovins adultes en Belgique suit une distribution normale (Laplace-Gauss), nous pourrions générer dans excel un échantillon du poids de 100 bovins pris au hasard en utilisant la fonction `LOI.NORMALE.INVERSE.N(proba;moyenne;deviation std)` considérant une moyenne de 750 kg et une variance de 2500 kg². Afin de produire la distribution empirique obtenue à partir d'échantillons tels que celui-ci, il faut regrouper les 100 poids dans des classes dont on calculera les fréquences. Créez un tableau dont la première colonne contiendra la limite supérieure de chaque classe. Par exemple, vous pourriez commencer à 550 pour la première classe et augmenter de 50 unités à chaque classe. Pourquoi commencer à 550 ? La deuxième colonne du tableau contiendra les effectifs non cumulés de chaque classe, à définir avec la fonction `FREQUENCE(plage données;plage lim classes)` en format matriciel

Cette distribution empirique est-elle bien une distribution normale ? Pour répondre à cette question, il faut réaliser un test de χ^2 de conformité. Vous devez donc générer les effectifs théoriques (ou attendus) en utilisant la fonction `LOI.NORMALE.N(X;moyenne;deviation std;cumulatif(vrai/faux))`, X étant la limite supérieure de chaque classe. En mettant vrai comme valeur pour "cumulatif", on obtient la probabilité cumulée de moins l'infini à la valeur X dans la 3^e colonne du tableau, il faudra donc enlever ce cumul par différence des classes cumulées. La somme de cette 4^e colonne vaudra alors 100%. Remarque : faites attention aux classes ouvertes, c'est-à-dire la première et la dernière classe lorsque vous « enlevez » le cumul des fréquences théoriques.

Le test de chi-carré peut alors s'effectuer, après avoir regroupé les classes dont les effectifs attendus sont inférieurs à 5 (prérequis à un test de chi-carré). Détaillez votre calcul du chi-carré, donnez-en le nombre de degrés de liberté et affichez la probabilité de s'écarter aussi fort ou plus fort des valeurs attendues sous l'hypothèse nulle. Quelle est votre conclusion ? Représentez graphiquement les deux distributions - théorique et empirique - sur le même graphique.

Réalisez ensuite ce même exercice sur R.