INTRODUCTION – TESTS DE COMPARAISON DE MOYENNES

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.

Lors des 5 séances de TP, vous serez amenés à manipuler le tableur électronique Microsoft Excel et le logiciel statistique open source R. L'introduction à l'utilisation de ces logiciels faisant partie de la matière de BMV1, prérequis au cours de BMV3, vous trouverez toutes les informations nécessaires pour vous refamiliariser avec ces outils sur la partie du site consacrée à la matière de BMV1. Vous trouverez notamment le lien vers le site ABC d'R (http://www.biostat.ulg.ac.be/pages/Site_r/ABC_R.html), vous expliquant concrètement le prise en main du logiciel pour atteindre les objectifs du cours et des TP.

Pour mémoire, le programme R (open source gratuit) est installé sur chaque ordinateur de la salle. Quand vous l'ouvrez, commencez par changer son répertoire de travail (ou répertoire courant) via le menu Fichier. Choisissez un répertoire qui vous est propre (Mes Documents) ou un support externe (clé USB). Tout au long des directives que vous lui donnerez, R travaillera avec ce répertoire choisi, que ce soit pour lire des fichiers (importer des données à analyser) ou pour écrire des fichiers (exporter des données ou des résultats d'analyse).

Dans R, l'aide sur les paramètres / arguments nécessaires aux fonctions à utiliser peut être obtenue de deux manières. Par exemple, pour la fonction *mean(…)* :

```
> help("mean")
> ?mean
```

Exercice 1

Pour illustrer la manière dont les tests de comparaison de moyennes fonctionnent, nous allons créer un exemple et utiliser le test de t qui permet de voir si deux échantillons proviennent de deux populations de mêmes moyennes. Comme on supposera en outre que les populations dont sont issues les deux échantillons sont normales et de mêmes variances, ça revient à dire que l'on teste si les deux populations sont identiques. La première étape consiste à générer deux séries de mesures (qu'on supposera prises sur des individus différents, càd. non pairées. Par ex., si on choisit 30 individus, 15 feraient partie du premier groupe et les 15 autres du second groupe) et qui vont constituer nos deux échantillons.

Dans Excel, employez la fonction *LOI.NORMALE.INVERSE(ALEA();μ;s)* pour générer deux séries de valeurs (les deux séries peuvent ou non avoir le même nombre de données, si vous choisissez d'échantillonner le même nombre de données dans chaque série, vos données seront qualifiées de "balancées"). On peut par exemple comparer le poids de bovins BBB mâle à 18 mois, selon le régime alimentaire fourni. Vous allez ainsi échantillonner deux groupes (Régime 1 vs Régime 2), et si vous souhaitez imposer une différence de poids entre les régimes, vous choisirez deux moyennes différentes entre les deux groupes. La variance pouvant rester identique car les bovins sont issus de la même population de départ. La moyenne et la variance sont laissées à votre appréciation, mais restez crédible dans le poids d'un bovin et dans la manière dont les poids varient autour de la moyenne (des poids négatifs risquent d'apparaître si vous optez pour une variance trop grande !). Pour information, dans une distribution normale, environ 95% des individus se retrouvent entre les limites $\mu-2*\sigma$ et $\mu+2*\sigma$. Afin de limiter le nombre de décimale dans les poids générés, vous pouvez utiliser la fonction *ARRONDI(…)* sur la formule qui génère les valeurs.

NB: avec la fonction ALEA() vos données échantillonnées changent à chaque frappe dans la feuille Excel. Si vous souhaitez travailler sur de données "figées", vous pouvez faire un copier-collage spécial des valeurs des données aléatoires et les coller juste à côté. Vous conserverez ainsi vos formules de départ.

Après avoir générer vos 2 colonnes d'individus, quel test statistique allez-vous mettre en œuvre pour comparer vos deux moyennes et pourquoi ? Pour cela, calculez les différents paramètres nécessaires pour chaque groupe comme si vous ne connaissiez pas les paramètres (choisis) de la population de départ, à savoir la moyenne et la variance (ou la déviation standard). Vous aurez donc besoin des paramètres estimés suivants pour le calcul de la statistique : nombre de données *NB(…)*, moyenne *MOYENNE(…)*, variance *VAR(…)* ainsi que la variance commune entre les deux groupes. Cette variance commune est une moyenne pondérée par les degrés de liberté de chaque groupe (nombre de données du groupe -1) de chaque variance estimée S^2 soit $(S_1^2*(n_1-1)+S_2^2*(n_2-1))/(n_1+n_2-2))$. N'oubliez de mettre des noms (étiquettes) dans les cellules pour les différents paramètres que vous calculez.

Un point important à noter est de formuler votre hypothèse nulle, à savoir **H0: µ1 = µ2**. Nous sommes dans le cas d'un **test bilatéral** càd. pas d'attente qu'un régime en particulier donne un poids moyen supérieur à l'autre. Nous regardons simplement si les deux régimes donnent des poids moyens significativement différents l'un de l'autre ou non.

Une fois la valeur de votre test statistique obtenue, il va falloir comparer celle-ci à une valeur de référence afin de prendre votre décision finale (accepter ou rejeter l'hypothèse nulle). Deux méthodes sont possibles:

- soit grâce à la table de la statistique appliquée, nous recherchons la limite théorique à ne pas dépasser pour accepter l'hypothèse nulle (*LOI.STUDENT.INVERSE(α;ddl)*)
- soit on calcul la probabilité, si l'hypothèse nulle est vraie, d'obtenir une telle valeur de notre statistique (*LOI.STUDENT(t;ddl;1/2)*)

Bien évidemment, l'acceptation ou le rejet de l'hypothèse nulle se fera toujours après avoir choisi un seuil limite (habituellement alpha = 5%).

Exercice 2

Vous venez d'appliquer un test statistique connu, un peu comme une recette de cuisine… Mais si vous ne connaissiez aucun test statistique et que vous souhaiteriez comparer les moyennes des vos deux groupes, comment pourriez-vous vous y prendre ?

"Simplement" en faisant un grand nombre de permutations aléatoires des données entre les groupes. En d'autres termes, vous pourriez réattribuer aléatoirement (au hasard) l'appartenance de chaque donnée à un groupe. Cela part de l'hypothèse (nulle) disant qu'il n'y a pas de différence entre les moyennes des groupes, et donc, dans cette optique n'importe quelle donnée pourrait appartenir à n'importe quel groupe. On s'attendrait donc une différence de moyennes ($\mu_1 - \mu_2$) nulle, mis à part suite aux fluctuations d'échantillonnage… ou bien si un effet était réellement présent (un régime donne réellement un poids moyen supérieur à l'autre!).

Mais comment procéder…?

Dans Excel, faites un copier-collage spécial de vos données si vous ne l'avez pas fait auparavant. Ensuite copier ces valeurs figées dans un nouveau classeur, dont la colonne A serait le groupe (Régime 1 vs Régime 2) et la colonne B, le poids. Dans cette disposition en colonne, chaque ligne représente donc un individu et chaque colonne est un paramètre (facteur ou variable) appartenant à l'individu. Dans notre situation, chaque individu aura deux paramètres, son groupe (appelé Régime) et son poids. Nous vous conseillons de mettre un en-tête à chaque colonne utilisée (càd. dans la ligne 1 de la feuille Excel. Par exemple, cellule A1 : Regime). Enregistrez ce nouveau fichier avec un nom

simple et au format CSV (Type de fichier : "CSV (séparateur: point-virgule)"), mais surtout dans le répertoire de travail du programme R que vous avez choisi (cf. Introduction de ce TP).

NB: quand vous attribuez un nom à une variable ou un fichier, essayez d'utiliser des noms simples, sans accents ni caractères spéciaux et sans espace entre les mots (utilisez l'underscore ou tiret bas pour remplacer l'espace).

Dans le programme R, importez votre fichier de données grâce à la fonction read.csv2(…) et nommez simplement cette importation de données (d par exemple).

Grâce à la fonction *t.test(Poids~Regime,data=d,var.equal=TRUE)*, vous obtiendrez normalement le même résultat que le test réalisé dans Excel.

Voici par exemple ce que vous pourriez obtenir dans le cas de deux échantillons de 10 bovins chacun :

```
Two Sample t-test

data:  Poids by Regime
t = -0.24556, df = 18, p-value = 0.8088
alternativehypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -29.62205  23.42205
sample estimates:
mean in group 1 mean in group 2
          591.4           594.5
```

Essayez de bien comprendre chacun ligne de ce package de résultats …

La fonction *sample(…)* va vous permettre d'échantillonner au hasard dans vos données. Si vous échantillonnez la totalité de vos données, *sample(…)* leur donnera un ordre aléatoire. Vous pouvez ainsi grâce à la fonction *data.frame(…)* récréer un nouveau dataset (avec un nouveau nom pour ne pas écraser l'original) et calculer les moyennes de chaque régime (fonction *mean(…)*). Toutefois, vous n'avez pas nécessairement besoin de recréer un nouveau dataset pour calculer la différence de moyennes entre le régime1 et le Régime2. En fonction de la taille d'échantillon que vous avez choisie, vous savez que les X premières valeurs correspondent au Régime1 et les autres au Régime2. Ce qui nous intéresse au final c'est cette différence entre les moyennes, et surtout de la comparer (en valeur absolue pour un test bilatéral) à la différence de moyennes des groupes de départ. Si vous réalisez par exemple 1000 fois cette comparaison, vous pourrez calculer le pourcentage de différences de moyennes « aléatoires » supérieures à la différence de moyennes des groupes de départ (originaux).

NB: pour faciliter certains calculs vous pouvez utiliser la fonction attach(…), permettant d'utiliser directement le nom des variables de votre dataset (Par exemple, à la place de devoir noter d$Regime, vous pourrez directement noter Regime dans vos formules).

Afin de réaliser vos 1000 comparaisons, utilisez une boucle *for(…){…}* !

Conseil: quand vous créez un mini-programme, ouvrez une fenêtre de script dans R (ou un fichier Bloc-Notes) et entrez vos lignes de commande dans cette fenêtre. Ensuite pour exécuter vos instructions dans R, faites un simple copier-coller de la fenêtre du script vers le programme R. Si vous avez une erreur dans une de vos lignes de commande, cela permettra de facilement faire les modifications (ATTENTION, n'oubliez pas d'enregistrer régulièrement ce script…).

Tout s'est bien exécuté ? Si c'est le cas, le pourcentage obtenu (nombre de différences de moyennes supérieures à la différence de départ divisé par le nombre total de comparaisons) doit avoisiner le pourcentage du test de t.

Exercice 3

Il existe d'autres méthodes pour « mélanger » vos données au travers des différents groupes, et cela de manière complètement aléatoire (à la place de la fonction *sample(…)*).

Trouvez une autre méthode utilisable à la fois dans R et dans Excel (TRUC : pensez à la fonction *ALEA()*).

A côté de ces méthodes à créer vous-même, vous pouvez utiliser dans R la fonction *perm.t.test(…)*, issue du package RVAideMemoire (à charger dans R). Cette fonction vous donnera la valeur du test de t ainsi que la probabilité issue des permutations (999 par défaut).

Exercice facultatif pour le brainstorming

Dans R, générez la distribution empirique de la statistique t pour 18 degrés de liberté et deux échantillons balancés. Echantillonnez dans R vos deux groupes et calculez un nombre de valeurs aléatoires de t (par ex. 1000) que vous regrouperez dans un tableau. A partir de ces données, établissez aussi le graphique empirique de la statistique t (quelle différence par rapport au graphique théorique ?).