UGA STA401

Projet : travail individuel sur un cas réel

Objectif de ce projet :

Apprendre à utiliser les méthodes de statistiques descriptives et inférentielles sur un ou deux échantillons de données réelles et savoir restituer ses résultats et leurs interprétations que l'on validera ensuite par des tests statistiques avec spécification du risque d'erreur dans les conclusions données.

Pour cette mise en oeuvre il faudra avoir collecté **deux échantillons de données continues exprimés dans une même unité** (à valeurs dans un même intervalle) correspondant à deux variables continues observées sur une même population (cas d'échantillons appariés) ou à une variable continue observée dans deux populations différentes (cas d'échantillons indépendants).

Exemples de données :

- Mon temps de trajet quotidien pour me rendre sur mon lieu de travail le matin (X) et le temps de trajet de retour le soir (Y). L'unité statistique est dans ce cas un jour ouvré et on relève deux temps (x_i, y_i) chaque jour ouvré. Les deux échantillons collectés sont ici dit appariés car les mesures prises le sont sur une même unité statistique. Ils sont du coup de même taille. La problématique pourrait-être : passé-je plus de temps à me déplacer pour me rendre sur mon lieu de travail qu'à en revenir ?

- Le pouls d'une personne mesuré le matin, puis le soir observés sur un échantillon de n personnes ou mesurés n jours consécutifs chez une même personne (dans le premier cas l'unité statistique est une personne et dans le second un jour). Là encore les échantillons sont appariés. La problématique pourrait-être : le pouls est-il en moyenne plus élévé le soir que le matin.

- Dans un logement partagé par deux étudiants A et B on relève le temps passé sous la douche de chacun d'entre eux (X le temps passé par A et Y celui passé par B). Si on relève X et Y les mêmes jours pour n jours et que la quantité d'eau chaude est limitée il y aura un lien entre X et Y, l'unité statistique est un jour et dans ce cas encore les échantillons sont appariés. La question pourrait-être : qui de A ou B consomme en moyenne le plus d'eau dans sa douche ou reste le plus longtemps sous la douche ?

- Si par contre dans l'exemple précédent on relève n_X temps de douche pour A et n_Y temps de douches pour B sans que ces relevés soient nécessairement pris les mêmes jours (n_X n'est dans ce cas pas nécessairement égal à n_Y) l'élément numéro i de l'échantillon de X est le temps de la ième douche de A tandis que dans l'échantillon

de Y, y_i celui de la ième douche de B. Ces deux échantillons sont dits indépendants car ils correspondent à la collecte d'une même variable (temps passé sous la douche) pour des unités statistiques différentes (les premières sont les douches prises par A et les seconde celles prises par B). Ces échantillons seront alors dits indépendants. On peut dans ce cas aussi envisager la même problématique que précédemment.

Echantillons indépendants ou appariés ?

Il n'est pas toujours évident d'identifier la bonne situation. Disons, pour fixer les idées, que dans le cas appariés (c. à d. qui vont par deux et non pas appareillés !) cela fait sens de calculer la différence entre x_i et y_i (à condition bien sur qu'ils soient exprimés dans la même unité) et qu'il y a un lien probable entre les deux valeurs. Dans le cas d'échantillons indépendants on collecte un indicateur dans deux populations différentes, donc pour des individus différents. Ce serait le cas si le premier échantillons était constitués des pouls de n_X valeurs de pouls de femmes et le second de n_Y valeurs de pouls d'hommes et on pourrait se demander si le sexe a un effet en moyenne sur le pouls. Pour faire l'acquisitions d'échantillons indépendants on peut collecter un indicateur quantitatif (le pouls) et un facteur à deux niveaux (variable qualitative ayant deux modalités possibles, dans l'exemple le sexe).

Méthodologie :

1. Proposer une problématique
2. Trouver ou collecter deux échantillons appariés ou indépendants de tailles comprises entre 15 et 20 (dans le cas appariés ils seront de même taille) pour y répondre.
3. Décrire chacun des deux échantillons et les comparer à l'aide de graphiques et de leurs résumés numériques (moyenne, ecart-type,quantiles...). Interpréter.
4. Modéliser et estimer les paramètres inconnus des lois posées sur les variables X et Y qui auraient pu produire les données x_i et y_i collectées (dans le cas de variables continues on utilisera les modèles gaussiens et ces paramètres sont les moyennes et variances des variables modèles X et Y)
5. Estimer et mettre en oeuvre des tests statistiques sur les paramètres qui permettront de confirmer ou d'infirmer les conjectures et interprétation proposées à l'issue de la phase descriptive.

Rendus attendus :

Un rapport au format .pdf de quatre pages (maximum) et une annexe contenant les données. Il sera constitué des parties suivantes :

1. **Introduction** : y décrire la question d'intérêt les variables utilisées pour y répondre et les échantillons collectés (en précisant leurs tailles unités utilisées,...)
2. **Analyse descriptive :** des résumés graphiques et numériques avec leurs interprétations. En conclusion de cette partie la problématique envisagée qui sera issue de l'analyse descriptive.

3. **Analyse inférentielle :** On modélisera les deux variables associées à chacun des échantillons par une variable normale (où seront bien précisées les notations utilisées pour décrire les paramètres inconnus de chacune des deux variables).

   (a) Estimation et Intervalles de confiance

   (b) Tests sur chaque échantillon

   (c) Tests de comparaison des deux échantillons

4. **Conclusions :** les informations et conclusions qui se dégagent de l'étude menée et la réponse à la problématique posée. Les éventuelles difficultés de collecte de données, les problèmes que cela a pu susciter et les moyens d'y remédier.

Ce projet sera à rendre en deux temps et à déposer dans la partie travaux sur Chamilo.

Calendrier :

- DM1 : rendu des parties 1 et 2 et de l'annexe contenant les données fin février.
- DM2 : rendu des parties 3 et 4 fin avril.