Introduction et définitions de base

La statistique est, de façon très informelle, l'étude de données numériques recueillies à partir d'un nombre de sujets. Nous donnerons plus tard des définitions plus précises, mais ceci conviendra pour commencer.

Définition 1. *Une étude statistique comprend trois étapes:*

1. *La collecte de données.*
2. *L'analyse descriptive, qui consiste à trouver des caractéristiques (généralement numériques, mais pas toujours) nous permettant de résumer les propriétés essentielles de nos résultats. Le calcul de la moyenne serait un exemple.*
3. *L'analyse inférentielle, qui consiste à tirer des conclusions à partir des données elles-mêmes ou des caractéristiques trouvées lors de l'analyse descriptive, ainsi que d'établir le risque d'erreur de nos conclusions.*

La première étape ne sera pas étudiée dans ce cours, la deuxième sera vue durant les trois premiers cours et la troisième durant les trois cours suivants.

Dans le cadre d'une étude statistique, la population étudiée est l'ensemble de tous les objets (personnes, animaux, produits, etc...) auxquels l'étude s'intéresse et les individus d'une étude sont les objets individuels étudiés.

Remarque. *Quoique les mots "population" et "individu" suggèrent que l'on parle d'études sur des êtres humains, ce ne sera pas nécessairement le cas. Les individus d'une étude pourraient être des animaux, des produits, des essais d'un même produit, etc... Dans le cadre de l'informatique, il s'agira souvent d'essais succéssifs d'un même algorithme.*

Définition 2. *La variable statistique sur laquelle porte une étude est la caractéristique à laquelle l'étude s'intéresse. Elle peut être quantitative si ses valeurs sont numériques (taille, notes, salaire, etc...) ou qualitative sinon (état civil, ethnie, domaine de travail, etc...). Une modalité d'une variable statistique est une valeur que cette variable peut prendre.*

Une variable quantitative est dite discrète si elle ne peut prendre qu'un nombre réduit de valeurs. Elle est dite continue si elle peut prendre toute valeur à l'intérieur d'un intervalle (fini ou infini).

Un échantillon est un groupe d'individus choisis parmi la population sur lesquels on mesure les caractéristiques qui nous intéressent. It est dit aléatoire si les individus ont été choisis au hasard.

Par convention, on utilisera des majuscules (tout particulièrement X, mais pas seulement) pour désigner une variable statistique et des minuscules pour désigner les valeurs pour un individu.

Exemple 1. *On fait une enquête sur la structure familiale, la taille et le revenu des ménages d'un quartier.*

La population totale est l'ensemble de tous les ménages du quartier, avec les individus étant chaque ménage. Puisqu'on ne peut faire un recensement en règle, on choisit au hasard un échantillon de 100 ménages.

Pour chaque ménage, on s'intéresse à trois variables: une variable qualitative - la structure du ménage, dont les modalités seraient: famille nucléaire, personne seule, famille monoparentale, etc... - une variable quantitative discrète - le nombre de personnes que le ménage comprend, ne pouvant être qu'un nombre entier positif - et une variable quantitative continue soit le revenu total du ménage qui pourrait être n'importe quelle nombre réel positif.

Dépouillement et organisation des données

Définition 3. *Le dépouillement des données consiste à regrouper les données recueillies en classes dont on compte les effectifs.*

La fréquence absolue d'une classe est le nombre d'individus de l'échantillon se retrouvant dans la classe en question.

La fréquence relative d'une classe est la proportion des individus de l'échantillon se retrouvant dans la classe en question.

Pour une variable quantitative, les limites d'une classe sont les valeurs la délimitant des autres classes, et le point milieu la valeur au centre de celle-ci.

Remarque. *Les classes sont parfois faciles à choisir, et parfois pas. Supposons, par exemple, que l'on s'interroge sur la santé des élèves d'une école et, qu'à cet effet, on recueille l'âge et la taille de tous les élèves.*

Il est naturel d'utiliser l'âge en soi, le nombre d'années, comme classe, mais qu'en est-il de la taille? Les données risquent de varier beaucoup plus, et il ne sera pas très fréquent que deux élèves aient exactement *la même taille...*

À cause de cela, on devra probablement utiliser des intervalles de tailles, par exemple ne pas compter le nombre d'élèves mesurant 103,5 cm (probablement 1) mais plutôt le nombre d'élèves mesurant entre 100 et 105 cm. Dans ce cas, 100 et 105 seraient les limites de la classe et 102,5 son point milieu.

Méthode 1. *Le choix des classes dépend de beaucoup de facteurs externes.* En l'absence de toute autre considérations, *on pourra utiliser la formule de Sturges:*

$$k \approx 1 + 3,322 \log_{10} n$$

Ou k, arrondi à l'entier le plus proche, sera le nombre de classes et n le nombre d'individus de l'échantillon. Dans le cas d'une variable quantitative, toutes les classes devraient avoir la même différence entre leurs limites.

Il faut insister sur le fait que cette formule n'est qu'indicative, en pratique il y a beaucoup d'autres choses à considérer.

Exemple 2. *Continuons l'exemple 1. La formule de Sturges nous conseille d'avoir:*

$$1 + 3,322 \log_{10} n = 1 + 3,322 \log_{10} 100 = 1 + 3,322 \cdot 2 = 7,644 \approx 8 = k$$

Il est évident que ceci n'est pas valable pour la structure familiale car le nombre de classes qu'on aura pour cette variable est fixé par l'étude. Ce n'est pas non plus utile pour la taille du ménage où il est plus utile de prendre la valeur en soi, quoique cela peut nous donner une indication à placer une seule catégorie pour les ménages comprenant 8 personne ou plus. Pour les revenus, toutefois, il serait raisonnable de séparer nos valeurs en huit intervalles.

Présentation graphique des données

Parlons maintenant de quelques formes classiques de présentation des données. Pour ne pas alourdir les notes de cours, nous ne présenterons ici que les définitions et les exemples seront donnés en classe.

Définition 4. *Un diagramme en bâtons présente sur l'axe des x le nom des classes, sur l'axe des y leur fréquence relative ou absolue, et la fréquence de chaque classe est représentée par la hauteur d'un rectangle au-dessus de chaque classe.*

Un histogramme est similaire à un diagramme en bâtons sauf que les rectangles sont collés ensemble et ce ne sont souvent pas les noms des classes qui sont indiquées sur l'axe des x mais les limites.

Un histogramme est préférable de loin lorsque les classes sont des intervalles numériques, dans les autres cas, les deux peuvent être utilisés de façon interchangeable. En général, on va utiliser un histogramme pour représenter une variable quantitative continue et un diagramme en bâtons pour représenter une variable quantitative discrète ou qualitative mais ceci n'est pas une règle absolue.

Définition 5. *Un polygone de fréquences présente sur l'axe des x le nom des classes ou les limites des classes, sur l'axe des y leur fréquence relative ou absolue, et la fréquence de chaque classe est représentée par un point au-dessus du point milieu de la classe. Les points sont reliés entre eux par une ligne.*

Si les classes sont placés dans un ordre quelconque (en particulier si la variable est quantitative), une courbe cumulative présente la somme des fréquences

des classes inférieures ou égales jusqu'à celle présentée.

On dessinera souvent le polygone de fréquences ou la courbe cumulative sur un histogramme.

Définition 6. *Un diagramme de secteurs présente la fréquence relative de certaines classes en colorant un secteur d'un disque.*

Définition 7. *Pour une variable quantitative qui change en fonction du temps, un graphe chronologique présente le temps le long de l'axe des x et la valeur sur l'axe des y, en faisant correspondre les deux avec des points reliés par une courbe.*

Exercices

Voici les résultats des 30 étudiants d'une classe à un examen:

45 89 53 94 52 87 47 56 60 97

54 94 51 67 52 97 94 50 54 95

54 75 94 55 58 59 98 82 55 94

1. Combien de classes la formule de Sturges vous suggère-t-elle d'avoir?
2. Choisissez des classes de taille appropriée étant donné la plus petite et plus grande note. Assurez-vous que 60 - une valeur plutôt important - soit l'une des limites.
3. Faites le dépouillement en fonction des classes choisies plus haut, puis reprèsentez les fréquences absolues par classe en utilisant un type de graphique approprié.
4. Représentez à l'aide de deux types de graphique appropriés la proportion de étudiants ayant passé ($\geq$ 60), coulé définitivement ($\leq$ 54) ou ayant droit à une recorrection (55-59).