Statistique

Nature des données

Contexte

Soit **x** une v.a. évaluée sur les objets d'un échantillon **E** et qui prends des valeurs dans un ensemble de référence **K**

$$x : E \rightarrow K : e \mapsto x(e)$$

où **E** est un sous-ensemble de la population **P**

Hypothèse: on suppose que P, E, x, K, sont bien définie en fonction des objectifs d'une étude statistique.

Nature des éléments de P, E

Les objets de P sont en lien avec le contexte de l'étude

Ça peut-être n'importe quoi: nombres, pays, chiens, personnes, etc

Ce ne sont pas des données!

Les objets de E sont de même nature que ceux de P, car $E \subset P$

Rappelons que l'ensemble P peut-être définie par

- extension: on explicite chaque élément
- compréhension: via une propriété commune aux éléments

Par exemple, considérons les pays du G7 comme population:

extension: P = {Allemagne, Canada, E-U, France, Italie, Japon, UK}

compréhension: P = { a est un pays et a est dans le G7 }

Population versus temps

Le contenu d'un ensemble a souvent une dépendence sur la variable "temps".

Par exemple, le G7 pourrait ne plus exister en 2020, ou encore l'Italie pourraît être remplacée par la Suisse, etc

Si l'étude porte sur une population à un moment précis, il faut le mettre dans la définition de la population

Par exemple: P = { a est un pays | a est dans le G7 au 2Q14 }

Si les objets changent dans le temps, on doit spécifier la période

Par exemple: Q = { symbole coté sur le TSX en 2014 }

Remarque: Q contient les symboles, pas les cotes, ces dernières sont des v.a. sur Q

Échantillon versus temps

L'échantillon peut aussi dépendre du temps

Par exemple:

- population de tous les joueurs de la LNH depuis le début
- échantillon formé des joueurs de l'équipe "Canadien de Mtl"

Si l'étude porte sur un moment précis, il faut le mettre dans la définition de l' échantillon. Par exemple:

{ les joueurs de l'équipe "Canadien de Mtl" au 31 décembre 2014 }

C'est l'étude statistique qui dicte le choix de la population, l'échantillon, les v.a. et la genre d'analyse à faire sur les valeurs prise pas les variables

Ensemble des valeurs d'un v.a.

Considérons une v.a. $\mathbf{x : E \rightarrow K}$

L'ensemble K peut-être fini ou infini, contenir des mots, des symboles, des nombres, etc

Exemples: { oui, non } , $\mathbb{N}$, $\mathbb{R}$, { toutes les chaînes de caractères UTF-8 } , { Harper, Trudeau } , [0,100] , { 1, 2, 3, 5, 8, 13 } , etc

Variable aléatoire dynamique

Il arrive que la valeur d'une v.a. sur un élément de l'échantillon change au cours du temps. On dit qu'elle est "dynamique", par opposition à "statique"

Une suite de valeurs d'une v.a. dynamique obtenues dans le temps s'appelle une "série temporelle"

Exemple: considérons la population:

$$Q = \{ \text{symboles coté sur le TSX durant l'année 2014} \}$$

et la v.a

$$\text{cote: } Q \rightarrow \mathbb{R}^+, \text{ qui associe a un symbole de Q sa cote sur le TSX}$$

C'est une série temporelle. En effet, si on considère le symbole IMG de Q, on peut voir les valeurs de la série pour une période récente via cote(IMG)

Pratique courante

Dans un rapport d'étude, on laisse généralement au lecteur le soin de déduire la population, l'échantillon, les v.a. et les ensembles des valeurs à partir de la mise en contexte

C'est souvent une source d'erreur et de mauvaise interprétation des résultats

Il convient d'être rigoureux et fournir un contexte complet et précis