M4202C - Statistique inférentielle
Cours 0 - Introduction

2021/2022 - A. Ridard

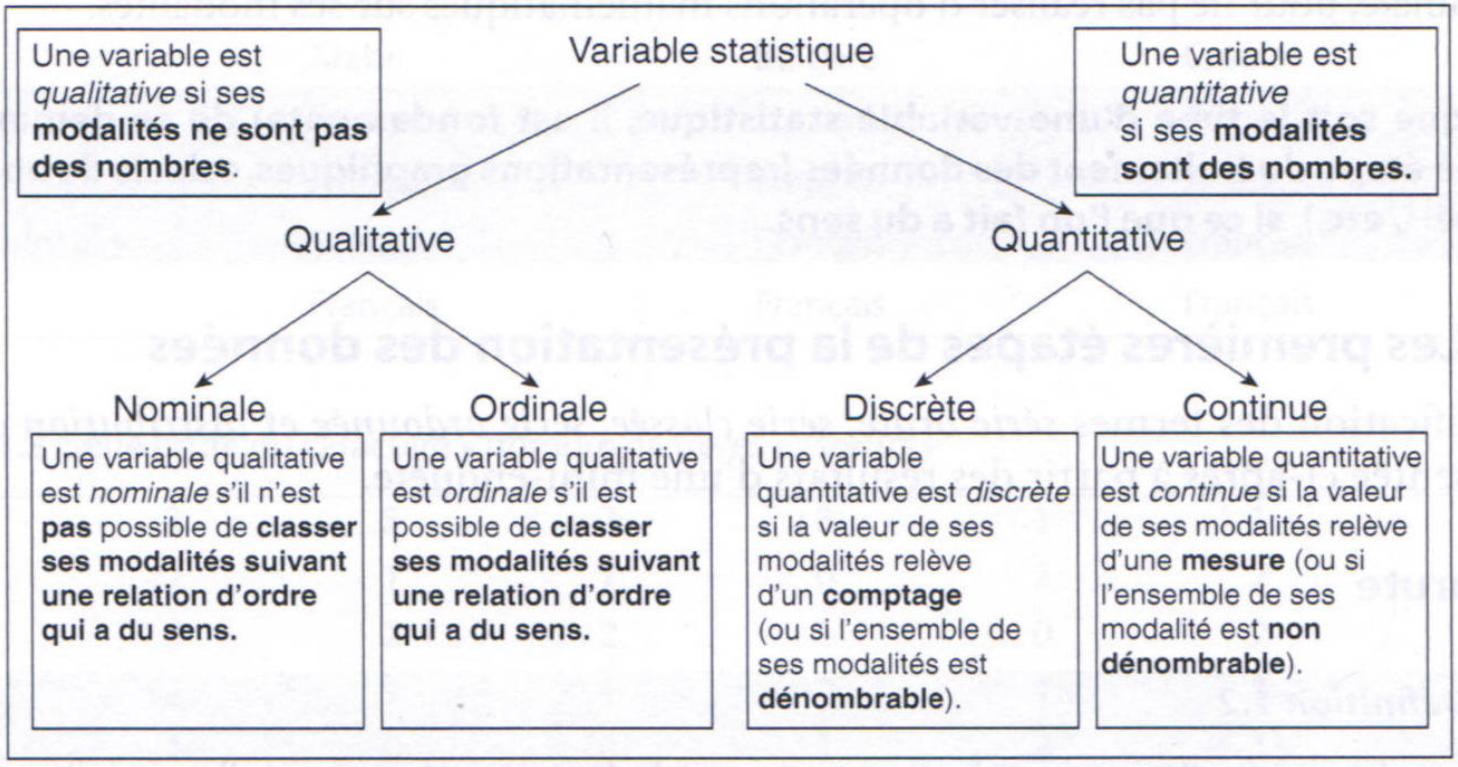
A propos de ce document

- Pour naviguer dans le document, vous pouvez utiliser :
  - le menu (en haut à gauche)
  - l'icône en dessous du logo IUT
  - les différents liens
- Pour signaler une erreur, vous pouvez envoyer un message à l'adresse suivante : anthony.ridard@univ-ubs.fr

Plan du cours

Si la Statistique descriptive consiste en l'étude d'une population toute entière d'individus selon un ou plusieurs caractères, la Statistique inférentielle permet de déduire des informations sur une population de taille N à partir d'un échantillon de taille n. Avant de préciser ce que l'on souhaite inférer, nous allons présenter différentes manières de prélever un échantillon.

Il est important de rappeler que chaque individu d'une population est caractérisé par un ou plusieurs caractères appelés aussi variables. On distingue deux types et quatre sous-types de variables.



Définition (variable aléatoire)

Une variable (statistique) est une application (au sens mathématique du terme) qui, à chaque individu, associe une valeur (numérique ou non).

Autrement dit, si l'on note ω un individu de la population et X la variable étudiée, alors $X(\omega) = x$ signifie que le caractère X a pour valeur x pour l'individu ω.

Si l'individu est choisi au hasard, la variable est dite aléatoire (v.a.).

Si, en Mathématiques, l'usage est plutôt d'appeler f, g ou h les applications, en Statistique celles-ci sont notées X, Y ou Z.
Les minuscules x, y ou z représentent alors les réalisations (valeurs) de ces variables (applications).
On fera donc bien la différence entre majuscules et minuscules pour éviter toute confusion entre applications et valeurs.

Dans toute la suite du cours, les v.a. considérées seront quantitatives.

Une variable aléatoire possède une loi de probabilité[a] qui régit son comportement.
Si la variable est discrète, la loi est définie par un diagramme en bâtons.
Si la variable est continue, le diagramme en bâtons est remplacé par une courbe de densité.

---

a. Si la Statistique fournit des informations sur une population à partir d'observations, les Probabilités fournissent des modèles théoriques pour étudier toute situation avec une part d'aléatoire.

Définition (n-échantillon aléatoire)

Un n-échantillon aléatoire est un n-uplet de v.a. $(X_1, ..., X_n)$ où X_i est le caractère X du i-ème individu choisi au hasard.

Définition (n-échantillon aléatoire simple)

Un n-échantillon aléatoire $(X_1, ..., X_n)$ est dit simple si les v.a. sont indépendantes.

Cela se produit si les individus sont choisis au hasard :

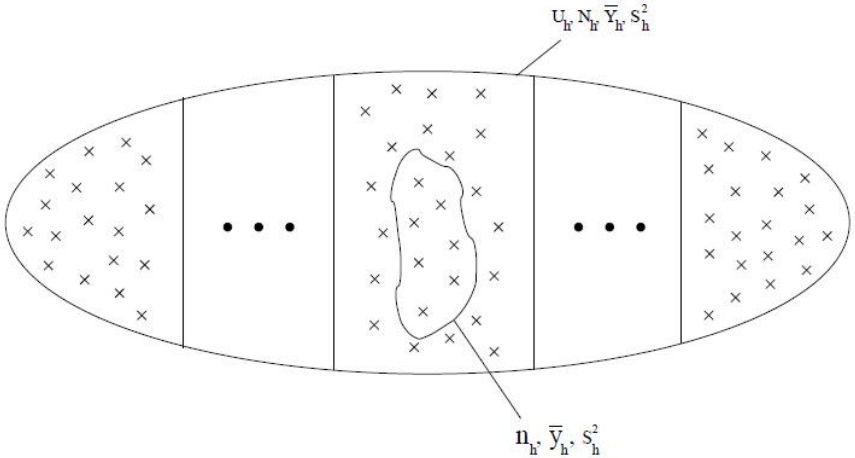
- soit avec remise
- soit sans remise (ou simultanément) à condition que le taux de sondage $\frac{n}{N}$ soit inférieur à 10%.

Le premier cas est théorique et le deuxième pratique.

Dans un sondage aléatoire simple, tous les échantillons d'une population de taille N sont possibles avec la même probabilité. On imagine que certains d'entre eux puissent s'avérer a priori indésirables.

Plus concrètement, dans l'étude du lancement d'un nouveau produit financier, on peut supposer des différences de comportement entre les "petits" et les "gros" clients de la banque. Il serait malencontreux que le hasard de l'échantillonnage conduise à n'interroger que les clients appartenant à une seule de ces catégories, ou simplement que l'échantillon soit trop déséquilibré en faveur de l'une d'elles. S'il existe dans la base de sondage une information auxiliaire permettant de distinguer, a priori, les catégories de "petits" et "gros" clients, on aura tout à gagner à utiliser cette information pour répartir l'échantillon dans chaque souspopulation.

C'est le principe de la stratification : découper la population en sous-ensembles appelés strates et réaliser un sondage aléatoire simple dans chacune d'elles.



Nous avons alors deux manières de choisir les n_h :

- allocation proportionnelle : $\forall h, \frac{n_h}{N_h} = \frac{n}{N}$
- allocation optimale [1] : $\forall h, \frac{n_h}{N_h} = n\frac{S_h}{\sum_h N_h S_h}$.

Dans toute la suite du cours, les échantillons aléatoires seront supposés simples.

1. On cherche la répartition de l'échantillon qui maximise la précision (et donc qui minimise la variance). Pour cela, on va augmenter les effectifs échantillonnés dans les strates où la variabilité est grande et diminuer les effectifs échantillonnés dans les strates homogènes.

On note X la v.a. étudiée.

On verra comment on peut inférer la loi de X, mais on s'intéressera avant tout aux informations suivantes :

- sa moyenne définie par :

$$m = E(X) = \begin{cases} \sum_i x_i P(X = x_i) & \text{si } X \text{ est discrète} \\ \int_{\mathbb{R}} x f(x)\, dx & \text{si } X \text{ est continue de densité } f \end{cases}$$

- sa variance définie par :

$$\sigma^2 = V(X) = E\left((X-m)^2\right) = \begin{cases} \sum_i (x_i - m)^2 P(X = x_i) & \text{si } X \text{ est discrète} \\ \int_{\mathbb{R}} (x-m)^2 f(x)\, dx & \text{si } X \text{ est continue de densité } f \end{cases}$$

- Rappelons les formules de Statistique descriptive pour les 3 types de données suivants :
  1. Les données brutes (individu par individu) de la forme :

$$x_1,\ x_2,\ \ldots,\ x_n$$

  2. Les données regroupées par valeur, dans le cas discret, de la forme :

x_1	x_2	...	x_p
n_1	n_2	...	n_p

  3. Les données regroupées par classe, dans le cas continu, de la forme :

$[e_1, e_2[$	$[e_2, e_3[$	...	$[e_p, e_{p+1}[$
n_1	n_2	...	n_p

- En notant $n = \sum_{i=1}^{p} n_i$ l'effectif total, $f_i = \frac{n_i}{n}$ la fréquence associée à n_i et $c_i = \frac{e_i + e_{i+1}}{2}$ le centre de la classe $[e_i, e_{i+1}[$, on a :

	1	2	3
$\bar{x}$	$\frac{1}{n}\sum_{i=1}^{n} x_i$	$\sum_{i=1}^{p} f_i x_i$	$\sum_{i=1}^{p} f_i c_i$
s^2	$\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$	$\sum_{i=1}^{p} f_i (x_i - \bar{x})^2$	$\sum_{i=1}^{p} f_i (c_i - \bar{x})^2$

- Les formules théoriques dans le cas discret sont à comparer avec celles de la colonne 2 !

On s'intéressera à la proportion p c'est à dire à la part des individus dans une population possédant un certain caractère.

p est aussi la moyenne de la v.a. de Bernoulli qui, à un individu, associe 1 s'il possède le caractère désiré et 0 sinon.
Les résultats sur p se déduiront donc de ceux sur m en prenant $X \sim \mathscr{B}(p)$.