

Préambule



I. Qu'est-ce que la statistique ?

Quelques citations amusantes glanées ici et là :

Edmond et Jules de GONCOURT :	« La statistique est la première des sciences inexactes. »
Alphonse ALLAIS :	« La statistique a démontré que la mortalité dans l'armée augmente sensiblement en temps de guerre. »
Mark TWAIN :	« Il existe trois sortes de mensonges : les mensonges, les sacrés mensonges et les statistiques »
Adolphe THIERS :	« ... L'art de préciser les choses que l'on ignore »
LAVELEYE :	« ... L'art de mentir mathématiquement »
MACAULEY :	« Les chiffres disent toujours ce que souhaite l'homme habile qui sait en jouer »
Louis ARMAND :	« Les statistiques , c'est comme le bikini, ça montre tout, mais ça cache l'essentiel »

Plus sérieusement :

ENCYCLOPEDIA UNIVERSALIS : « Le mot statistique désigne à la fois un ensemble de données d'observations et l'activité qui consiste dans leur recueil, leur traitement et leur interprétation »

PETIT LAROUSSE : (du latin status, état) « Ensemble de méthodes mathématiques qui, à partir du recueil et de l'analyse de données réelles, permettent l'élaboration de modèles probabilistes autorisant les prévisions »

Daniel SCHWARTZ : « La statistique est un mode de pensée permettant de recueillir, de traiter et d'interpréter les données qu'on rencontre dans divers domaines, et tout particulièrement dans les sciences de la vie, du fait que ces données présentent une caractéristique essentielle : la variabilité. »

La variabilité est un concept fondamental en statistique : des individus apparemment semblables peuvent prendre des valeurs différentes.

Quelques exemples :

- Le nombre de quartiers dans le fruit du coquelicot : le plus souvent égal à 13, il peut varier de 6 à 20 !
- Le temps de latence d'une maladie virale du tabac (mosaïque) : il varie de 13 à 34 jours.
- En cours de fabrication, la longueur d'un profilé aluminium n'est jamais parfaitement constante.

...

Pour l'essentiel, l'analyse statistique est **une étude de la variabilité**.

2. Statistique et probabilités

La **théorie des probabilités** modélise des phénomènes où le « hasard » intervient. Comme toute théorie mathématique, elle est basée sur une axiomatique et se développe de façon autonome par rapport à la réalité physique.

La **statistique** repose, elle, sur l'observation de phénomènes.

Quels liens entre la statistique et les probabilités ?

G. Saporta en voit schématiquement trois.

- Les données observées sont souvent imprécises ou entachées d'erreur. La théorie des probabilités permet de représenter les déviations entre vraies valeurs et valeurs observées comme des variables aléatoires.
- On constate souvent que la répartition statistique d'une variable au sein d'une population est voisine de modèles mathématiques proposés par la théorie des probabilités (lois de probabilité).
- Les échantillons d'individus observés sont la plupart du temps tirés au hasard dans la population, ceci pour assurer mathématiquement leur représentativité : si le tirage est fait de manière équiprobable chaque individu de la population a une probabilité constante et bien définie d'appartenir à l'échantillon. Les caractéristiques observées sur l'échantillon deviennent, grâce à ce tirage au sort, des variables aléatoires et le calcul des probabilités permet d'étudier leurs répartitions.

Dans les deux premiers cas, la théorie des probabilités propose des modèles (simplificateurs mais peut-être contestables), du comportement d'un phénomène (par exemple la durée de vie X d'un composant électronique suit une loi exponentielle, c'est à dire $P(X > x) = \exp(-\lambda x)$).

Dans le dernier cas, la théorie des probabilités fournit des théorèmes si le processus d'échantillonnage est respecté : ainsi le **théorème central limite** permet d'établir que la moyenne \bar{X} d'une variable numérique mesurée sur n individus s'écarte de la moyenne m de la population entière selon une loi approximativement gaussienne.

Le **calcul des probabilités** est donc un des outils essentiels de la statistique pour pouvoir extrapoler à la population les résultats constatés sur l'échantillon.

3. La démarche statistique

On distingue deux grands aspects :

- L'aspect exploratoire : **statistique descriptive**

La statistique descriptive a pour objectif de synthétiser, résumer, structurer l'information contenue dans les données concernant un phénomène étudié. On utilise des représentations graphiques ou des tableaux et on calcule quelques indicateurs. Le rôle des modèles probabilistes est quasiment inexistant.

- L'aspect décisionnel : statistique inférentielle

Dans son document, **La statistique au collège**, Y. Olivier décrit la statistique inférentielle de la façon suivante :

En général, les ensembles d'observation correspondent à des échantillons liés au hasard (on dit aussi présentant un caractère aléatoire) et l'on essaie de modéliser le phénomène à l'aide de modèles probabilistes (on s'appuie sur certaines lois de probabilités classiques). Cela permettra sinon des prévisions tout du moins des présomptions qui sont précieuses dans l'étude de certains faits (sociaux, économiques ou industriels). L'**inférence** consiste donc à étudier les propriétés d'un échantillon représentatif d'un ensemble plus vaste et à généraliser ces propriétés en se souciant des questions suivantes : les faits étudiés sont-ils

significatifs et sont-ils caractéristiques de propriétés plus générales ? Elles permettent de prendre de «bonnes» décisions malgré la présence d'incertitudes comme dans la recherche de qualité de produit fabriqué ou comme dans les analyses en laboratoires pharmaceutiques par exemple.

La démarche du statisticien peut être illustrée par le schéma suivant extrait de la brochure Inter-IREM « *Liaison Collège-Secondaire* ».

