# Introduction

Pourquoi les statistiques?

*26 août 2019*

## Contents



## Quels types de questions font appel aux statistiques?

Les méthodes statistiques qui seront présentées dans ce cours sont utilisées dans l'ensemble des sciences naturelles et sociales. Il existe toutefois plusieurs raisons pour lesquelles ces méthodes occupent une place prépondérante dans certaines disciplines, comme l'écologie.

- Nous étudions des systèmes complexes, composées de nombreux types d'entités ou individus qui interagissent. Un changement donné peut engendrer une chaîne d'effets, des boucles de rétroaction, etc. Isoler l'effet d'une variable représente donc un défi important.
- Les individus ne sont pas identiques mais varient à plusieurs niveaux.
- Notre capacité d'observation du système est limitée, tant pour le nombre de variables mesurées, que le nombre d'individus échantillonnées, ainsi que par la précision des observations elles-mêmes.

## La méthode scientifique

La méthode scientifique est souvent présentée par un diagramme comme celui ci-dessous. Sur la base d'observations passées, des chercheurs émettent une *hypothèse* sur le fonctionnement d'un système. De cette hypothèse, on déduit certaines *prédictions*, qui sont comparées au résultat d'une *expérience* conçue spécifiquement pour tester l'hypothèse. Le résultat mène soit au rejet de l'hypothèse, soit à son acceptation provisoire (jusqu'à ce qu'une hypothèse compétitrice soit proposée sur la base de nouvelles observations ou résultats d'expérience).

En réalité, chaque étude scientifique ne suit pas ce schéma en entier. Différentes équipes peuvent réaliser les nouvelles observations, formuler de nouvelles théories ou hypothèses et concevoir les expériences visant à tester ces hypothèses. Pour certains systèmes, il n'est pas éthique ou pratique de réaliser des expériences contrôlées, donc les hypothèses doivent être testées à partir d'observations.

Il existe aussi plusieurs questions d'intérêt scientifique qui ne sont pas des tests d'hypothèse. Par exemple:

- Combien y a-t-il d'espèces d'oiseaux uniques à cette région?
- Quelle est l'aire de répartition du pin gris? Comment sera-t-elle modifiée par les changements climatiques au 21e siècle?

## Rôle des statistiques

Les quatre buts suivants résument la plupart des applications des statistiques:

- *Décrire* les caractéristiques sommaires d'un ensemble de données.
- À partir de mesures prises sur un échantillon d'individus, *estimer* les caractéristiques d'une variable, ou d'une relation entre variables, au niveau de la population.
- *Tester* une hypothèse portant sur ces variables ou relations entre variables.
- *Prédire* la valeur d'une variable pour un nouvel individu hors de l'échantillon.

**Exemples**

- Quel est le diamètre moyen des arbres mesurés dans une parcelle? (description, puisqu'on a la mesure pour tous les arbres)
- Combien y a-t-il d'orignaux dans le Parc de la Vérendrye? (estimation, puisqu'on ne les a pas tous observés)
  - La population est-elle en croissance par rapport à l'année précédente? (test d'hypothèse)
  - Quelle sera la population dans 10 ans? (prédiction)

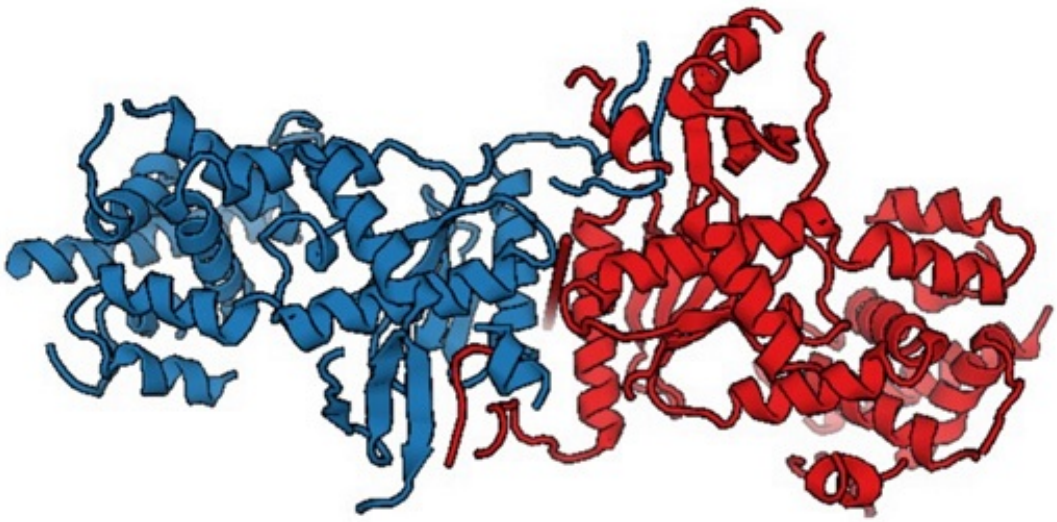
Les trois dernières tâches (estimation, test d'hypothèse et prédiction) demandent de généraliser les connaissances acquises à partir d'un nombre limité d'observations. Cela requiert un *modèle* du système étudié.

## Qu'est-ce qu'un modèle?

La modélisation est souvent associée aux mathématiques, mais l'idée de modèle ne se limite pas à une liste d'équations.

Par exemple, une carte routière constitue un modèle du réseau routier d'une région. Elle met l'accent sur des informations jugées prioritaires – routes principales et secondaires, distances entre les villes – mais ne peut représenter l'intégralité du réseau faute d'espace. La prolifération des appareils de navigation GPS nous permet maintenant d'utiliser des modèles plus complexes et plus informatifs du même réseau.


Sainte-Agathe-des-Monts
Val-David
Sainte-Adèle
Saint-Hippolyte
Saint-Calixte
Saint-Lin–Laurentides
Prévost
Lafontaine
Sainte-Sophie
Saint-Jérôme
Bellefeuille
Saint-Antoine
La Plaine
Sainte-Anne-des-Plaines
Saint-Colomban
Brownsburg-Chatham
Lachute
Mirabel
Blainville
Sainte-Thérèse
LAVAL
Boisbriand
Saint-Eustache
Saint-Joseph-du-Lac
Pierrefonds
Dollard-Des Ormeaux
Saint-Laurent
Rigaud



```
  0 MAVPGVGLLT RLNLCARRRT RVQRPIVRLL SCPGTVAKDL RRDEQPSGSV ETGFEDKIPK
 60 RRFSEMQNER REQAQRTVLI HCPEKISENK FLKYLSQFGP INNHFFYESF GLYAVVEFCQ
120 KESIGSLQNG THTPSTAMET AIPFRSRFFN LKLKNQTSER SRVRSSNQLP RSNKQLFELL
180 CYAESIDDQL NTLLKEFQLT EENTKLRYLT CSLIEDMAAA YFPDCIVRPF GSSVNTFGKL
240 GCDLDMFLDL DETRNLSAHK ISGNFLMEFQ VKNVPSERIA TQKILSVLGE CLDHFGPGCV
300 GVQKILNARC PLVRFSHQAS GFQCDLTTNN RIALTSSELL YIYGALDSRV RALVFSVRCW
360 ARAHSLTSSI PGAWITNFSL TMMVIFFLQR RSPPILPTLD SLKTLADAED KCVIEGNNCT
420 FVRDLSRIKP SQNTETLELL LKEFFEYFGN FAFDKNSINI RQGREQNKPD SSPLYIQNPF
480 ETSLNISKNV SQSQLQKFVD LARESAWILQ QEDTDRPSIS SNRPWGLVSL LLPSAPNRKS
540 FTKKKSNKFA IETVKNLLES LKGNRTENFT KTSGKRTIST QT
```

Comparons maintenant deux modèles d'une protéine dans l'image ci-dessus: en bas, une séquence d'acides aminés, chacun représentés par une lettre; en haut, une représentation de la structure 3D de la protéine. Chacune des deux représentations inclut une information absente de l'autre.

Finalement, voici un modèle mathématique bien connu en écologie. Le modèle de Lotka-Volterra représente l'évolution du nombre de proies ($x$) et de prédateurs ($y$) en fonction de quatre constantes: taux de reproduction des proies, taux d'attaque, efficacité de conversion et mortalité des prédateurs.

$$x_{t+1} = ax_t - bx_t y_t$$
$$y_{t+1} = -cy_t + dx_t y_t$$

Qu'est-ce que ces modèles ont en commun? Ils constituent des *représentations abstraites et simplifiées* d'un système. Au sens strict, tous les modèles sont inexacts. Cependant, il s'agit de concevoir le modèle le plus simple possible qui retient les caractéristiques essentielles du système pour un problème donné, tout en ignorant les détails qui sont superflus. On parle ici d'un principe de *parcimonie*.

Au sens mathématique, un modèle inclut des équations reliant les attributs observables, ou *variables*, des entités étudiées. En statistique, une *variable aléatoire* varie d'une observation à l'autre pour des raisons qui ne sont pas parfaitement connues. Ainsi, un modèle statistique associe à chaque variable aléatoire une *distribution* représentant la probabilité des différentes valeurs possibles.

## Aperçu du contenu du cours

- Semaine 1: Statistiques descriptives et visualisations graphiques.
- Semaine 2: Modèles statistiques, estimation de paramètres.
- Semaine 3: Méthodes d'échantillonnage et plans d'expériences.
- Reste du cours en trois parties:
  - Introduction aux tests d'hypothèses, dans le contexte de la comparaison des effets de deux ou plusieurs traitements. (3 semaines)
  - Modèles de régression: expliquer la variation d'une variable réponse à partir de prédicteurs numériques ou catégoriques. (6 semaines)
  - Analyses multivariées: lorsque la réponse à expliquer est composée de plusieurs variables. (2 semaines)