Introduction

The role of statistics

August 26, 2019

Contents

What types of questions require statistics?

The statistical methods that will be presented in this course are used throughout the natural and social sciences. There are, however, several reasons why these methods occupy a prominent place in certain disciplines, such as ecology.

- We study complex systems, composed of many types of entities or individuals that interact. A given change can result in a chain of effects, feedback loops, and so on. Isolating the effect of a variable is therefore a significant challenge.
- Individuals are not identical but vary on many levels.
- Our ability to observe the system is limited, both in the number of variables measured, the number of individuals sampled, and the accuracy of the observations themselves.

The scientific method

The scientific method is often presented by a diagram like the one below. On the basis of past observations, researchers pose a *hypothesis* on the functioning of a system. From this hypothesis, we deduce certain *predictions*, which are compared to the result of an *experiment* designed specifically to test the hypothesis. The result leads either to the rejection of the hypothesis or to its provisional acceptance (until a competing hypothesis is proposed on the basis of new observations or results of an experiment).

In fact, every scientific study does not follow this whole pattern. Different teams can make new observations, formulate new theories or hypotheses, and design experiments to test these hypotheses. For some systems, it is not ethical or practical to perform controlled experiments, so hypotheses must be tested from observations.

There are also several types of scientific questions that are not hypothesis tests. For example:

- How many species of birds are unique to this region?
- What is the distribution range of jack pine? How will it be changed by climate change in the 21st century?

Role of statistics

The following four goals summarize most statistical applications:

- *Describe* and summarize the characteristics of a dataset.
- From measurements taken on a sample of individuals, *estimate* the characteristics of a variable, or of a relationship between variables, at the population level.
- *Test* a hypothesis concerning these variables or relations between variables.
- *Predict* the value of a variable for a new individual out of the sample.

Examples

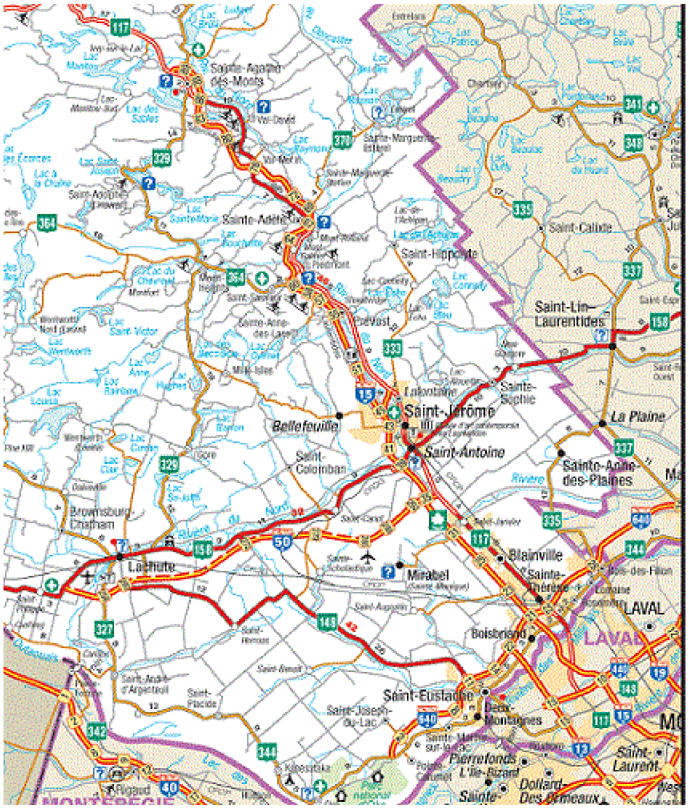
- What is the average diameter of the trees measured in a plot? (description, since we have measured all trees)
- How many moose are there in the Parc de la Vérendrye? (estimate, since we have not seen them all)
  - Is the population growing compared to the previous year? (hypothesis test)
  - What will be the population in 10 years? (prediction)

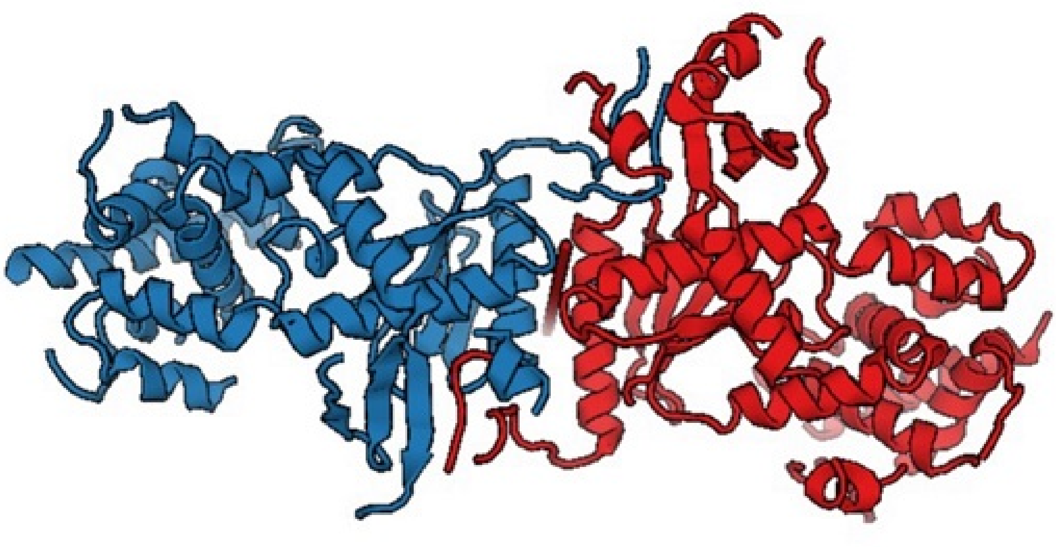
The last three tasks (estimation, hypothesis testing and prediction) require the generalization of knowledge acquired from a limited number of observations. This requires a *model* of the studied system.

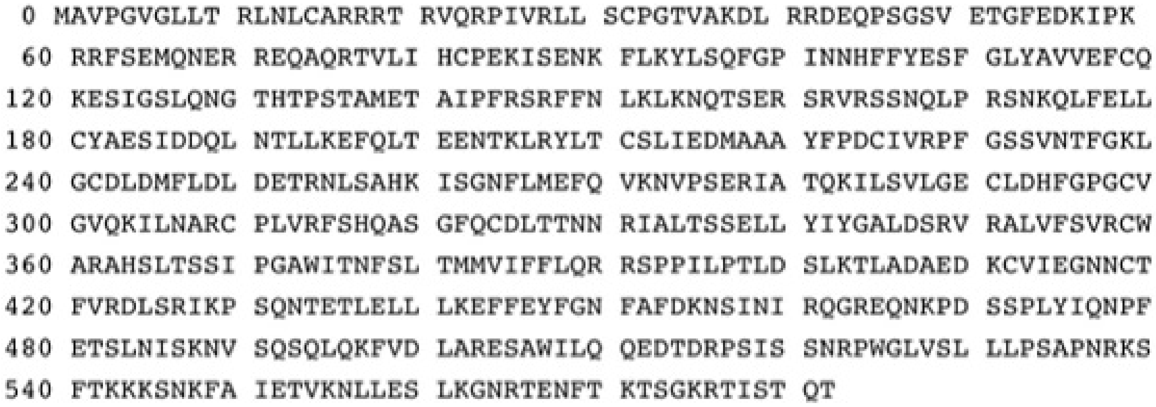
What is a model?

Modeling is often associated with mathematics, but the idea of a model is not limited to a list of equations.

For example, a road map is a model of the road network of a region. It focuses on information deemed to be important - main and secondary roads, distances between cities - but cannot represent the entire network due to space limitations. The proliferation of GPS navigation devices now allows us to use more complex and informative models of the same network.


Sainte-Agathe-des-Monts
Val-David
Val-Morin
Sainte-Marguerite-Estérel
Sainte-Marguerite-Station
Saint-Adolphe-d'Howard
Sainte-Adèle
Mont-Rolland
Piedmont
Saint-Hippolyte
Saint-Sauveur
Sainte-Anne-des-Lacs
Prévost
Mille-Isles
Chertsey
Saint-Calixte
Saint-Lin–Laurentides
Lafontaine
Saint-Jérôme
Sainte-Sophie
Bellefeuille
Saint-Antoine
Saint-Colomban
La Plaine
Sainte-Anne-des-Plaines
Brownsburg-Chatham
Lachute
Mirabel
Blainville
Sainte-Thérèse
Bois-des-Filion
LAVAL
Boisbriand
Saint-Eustache
Deux-Montagnes
Saint-Joseph-du-Lac
Saint-André-d'Argenteuil
Saint-Placide
Pierrefonds
L'Île-Bizard
Dollard-Des Ormeaux
Saint-Laurent
Rigaud
MONTÉRÉGIE
Lac Saint-Victor
Lac Wentworth
Lac Louisa
Lac des Sables
Lac Connelly
117
329
364
370
335
337
341
348
158
333
327
342
344
148
15
50
40
640
13
19





```
  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
```

Let's now compare two models of a protein in the image above: at the bottom, an amino acid sequence, each represented by a letter; at the top, a representation of the 3D structure of the protein. Each of the two representations includes information that is missing from the other.

Finally, here is a well-known mathematical model in ecology. The Lotka-Volterra model represents the evolution of the number of prey (x) and predators (y) according to four constants: prey reproduction rate, attack rate, conversion efficiency, and mortality of predators.

$$x_{t+1} = ax_t - bx_ty_t$$
$$y_{t+1} = -cy_t + dx_ty_t$$

What do these models have in common? They constitute *abstract and simplified representations* of a system. Strictly speaking, all models are inaccurate. However, the goal is often to design the simplest possible model that retains the essential characteristics of the system for a given problem, while ignoring the details that are superfluous. We speak here of a principle of *parsimony*.

In the mathematical sense, a model includes equations linking the observable attributes, or *variables*, of the entities studied. In statistics, a *random variable* varies from one observation to another for reasons that are not perfectly known. Thus, a statistical model associates with each random variable a *distribution* representing the probability of the different possible values.

Overview of course content

- Week 1: Descriptive statistics and graphics.
- Week 2: Statistical models, estimation of parameters.
- Week 3: Sampling methods and design of experiments.
- Rest of the course in three parts:
  - Introduction to hypothesis testing, in the context of comparing the effects of two or more treatments. (3 weeks)
  - Regression models: Explain the variation of a response variable from numerical or categorical predictors. (6 weeks)
  - Multivariate analyses: when the response to explain is composed of several variables. (2 weeks)