Statistique

I – Introduction

Richard Wilson

(richard.wilson@upmc.fr)

LATMOS/IPSL
Université Pierre & Marie Curie

12 septembre 2016

Sommaire

Introduction I

Définitions

- Le mot **statistique** dérive du latin **status** qui signifie «état».
- Le mot «statistique» est employé aujourd'hui à double sens :
  - **LA** statistique recouvre un ensemble de méthodes mathématiques dédiées à l'analyse et à l'interprétation de données.
  - **UNE** statistique est un nombre calculé à partir d'observations. Par exemple :
    - √ la moyenne ou la fréquence d'occurrence d'un évènement sont des statistiques.
    - √ On parle fréquemment des statistiques du chômage

Introduction II

- Introduit au dix-huitième siècle (vers 1785) par l'économiste allemeand Gottfried Achenwall, le mot «Statistik» est dérivé de l'italien «statista» (« homme d'État »),
- la statistique représente pour cet auteur l'ensemble des connaissances que doit posséder un homme d'État (état des ressources, évolution des prix, démographie).
- Si le mot est récent, la pratique est aussi ancienne que les sociétés humaines.
  - √ Les premiers textes écrits retrouvés (Sumériens) étaient des recensements du bétail et des informations sur les prix en cours.
  - √ On a ainsi trace de recensements des récoltes en Chine au XXIII^{e} siècle av. J.C. (empereur Yao), et de la population en Égypte au XVIII^{e} siècle av. J.C. (pharaon Amasis) (`http://www.statistix.fr/IMG/pdf/Une_approche_historique_de_la_statistique_v3.pdf`)

Introduction III

- La statistique mathématique moderne, basée sur le calcul des probabilité, est née au XIX^{e} siècle suite aux travaux de Laplace, Gauss et Moivre.
- La statistique connaît un développement fulgurant au tournant du vingtième siècle en Angleterre sous l'impulsion décisive de Francis Galton, Karl et Egon Pearson (le père et le fils), Ronald Fisher et William Gosset.

William Gosset en 1908

Introduction IV

- Au vingtième siècle, les techniques et méthodes statistiques se sont propagées et imposées dans tous les domaines scientifiques :
  - √ sciences "dures" (physique, astronomie, géophysique),
  - √ sciences du vivant (biologie, agronomie, médecine),
  - √ sciences économiques et sociales (sociologie, psychologie, histoire, économétrie),
- et bien au delà des sciences
  - √ publicité, marketing ;
  - √ production : sécurité (pannes, risques), contrôle de qualité ;
  - √ politique.
- Les méthodes statistiques sont devenues des outils d'aide à la décision (investissement financier, prévention des risques,...) voire des moyens opérationnels de gestion (files d'attente, circulation automobile, distribution d'électricité,...).
- Les progrès récents de l'informatique ont permis le développement de méthodes statistiques nouvelles, permettant la manipulation de très grandes quantités de données (exploration/fouille de données, data mining).

Introduction V

Il y a deux grandes branches de l'analyse statistique :

1. la **statistique descriptive** consistant au traitement, à la classification et à mise en forme des données. Une population sera résumée en quelques statistiques, un nuage de points sera projeté sur un système d'axes mettant en évidence des sous-groupes, etc...
2. la **statistique inférentielle** dont l'objet est la caractérisation des propriétés d'une population à partir d'échantillons de celle-ci. Les méthodes de la statistique inférentielle repose sur la théorie des probabilités.

La statistique descriptive

Le but de la statistique **descriptive** (ou **exploratoire**) est de structurer les données issues de l'échantillonnage afin d'en extraire des informations pertinentes.
C'est **l'analyse des données** dont les principales méthodes sont :

- √ Analyse univariée
  - L'estimation des quelques statistiques résumant une une collection de données (moyenne, médiale, dispersion,...) ;
- √ Analyse multivariée
  - Les tableau de contingence ;
  - Les graphiques de dispersion (scatter plots) visant à synthétiser les propriétés d'un échantillon ou d'une population ;
  - La mesure quantitative d'un lien de dépendance entre variables (corrélation, covariance)
  - Les méthodes factorielles ayant pour but de réduire le nombre de variables à l'aide d'un petit nombre de facteurs, comme par exemple l'analyse en composantes principales.

Le calcul des probabilité ne joue ici aucun rôle.

La statistique inférentielle

Le propos de la **statistique inférentielle** est de déduire les propriétés d'un ensemble, on dira (pour des raisons historiques) d'une population, à partir de la connaissance d'échantillons de cet ensemble.

L'échantillon peut résulter d'une série de mesures, d'un prélèvement aléatoire ou d'un sondage.

- série temporelle de mesures météorologiques au parc Montsouris (1872–)
- prélèvement aléatoire de pièces manufacturées (contrôle de qualité)
- sondage d'opinion.

La théorie des probabilités joue ici un rôle central.

Voici quelques exemples :

Exemple 1

- **Estimation d'une grandeur à partir d'une série de mesures**.

  Une grandeur physique est mesurée n fois. À cause des imprécisions de mesure ou à cause de la variabilité de facteurs ayant une influence sur la quantité mesurée, on recueille une suite de résultats $x_1, x_2, \ldots, x_n$, à priori tous différents.

  La mesure est une variable aléatoire (v.a.), X. L'objectif est d'estimer une valeur pertinente pour X à partir des $x_1, \ldots, x_n$. Une estimation intuitive pour $\overline{X}$ est la moyenne arithmétique des x_i, i.e. $\boxed{\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i}$ (statistique descriptive).

  Un résultat important de la théorie des probabilités – la loi des grands nombres – montre que la moyenne empirique $\overline{X}$ est une bonne **estimation** de X. Cependant, une autre série de mesures aurait eu, à priori, une autre moyenne empirique. La moyenne $\overline{X}$ est donc une variable aléatoire dont on cherchera à préciser les propriétés (statistique inférentielle). On cherchera en particulier un intervalle de confiance : $[\overline{X} - \Delta\overline{X}; \overline{X} + \Delta\overline{X}]$.

Exemple 2

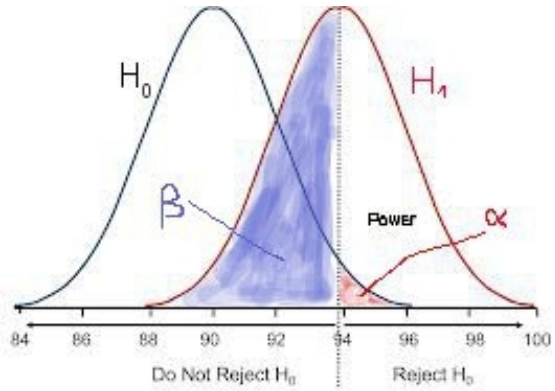
- **Test d'hypothèse**. Il s'agit ici de confirmer ou d'infirmer une hypothèse à partir de l'observation d'un échantillon ou d'une série de mesures.

  Considérons par exemple un processus dépendant du temps. On cherche à savoir si le processus est stationnaire, c'est à dire si sa valeur moyenne dépend ou non de l'intervalle de temps considéré.

  On estimera donc des moyennes dans des intervalles de temps consécutifs. On évaluera la pertinence de l'hypothèse de stationnarité en comparant la distribution des moyennes observées avec une distribution théorique attendue.

  Un test d'hypothèse permettra de prendre une décision (stationnarité ou non) avec un risque quantifié de se tromper.

Exemple 2



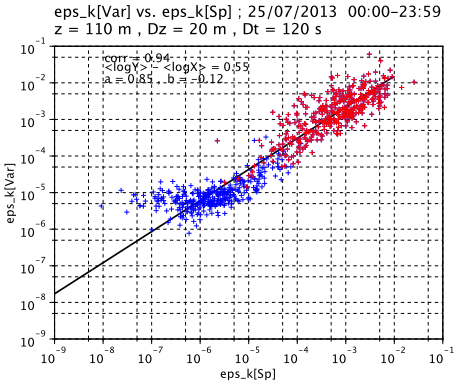
Principe du test d'hypothèse

Exemple

- **La modélisation statistique**. On cherche à établir une relation entre deux variables en estimant une fonction reliant les variations de l'une en fonction des variations de l'autre.

  Le plus souvent on recherche une relation linéaire via une droite de régression.

Exemple 3



Régression linéaire : $\log Y = a \log X + b$,

i.e. $Y = KX^a \qquad (K = 10^n)$.

Objectifs & évaluation

L'objectif de ce cours est double :

- Introduire des méthodes de statistique descriptive applicable à des populations multivariées (caractérisées par plusieurs variables)
- Donner quelques méthodes et outils informatiques permettant d'aborder des problèmes concrets. Le logiciel python sera utilisé.

Cet enseignement comprendra

- des cours magistraux.
- des séances de travail sur ordinateur illustrant les concepts du cours : traitement de données sous python,
- du travail personnel (écriture de scripts).

Évaluation

- Travail personnel à rendre (25%).
- Examen final sous forme d'un rapport de quelques pages et d'un exposé oral portant sur l'étude statistique d'un jeu de données (de votre choix si vous le souhaitez) (75%).